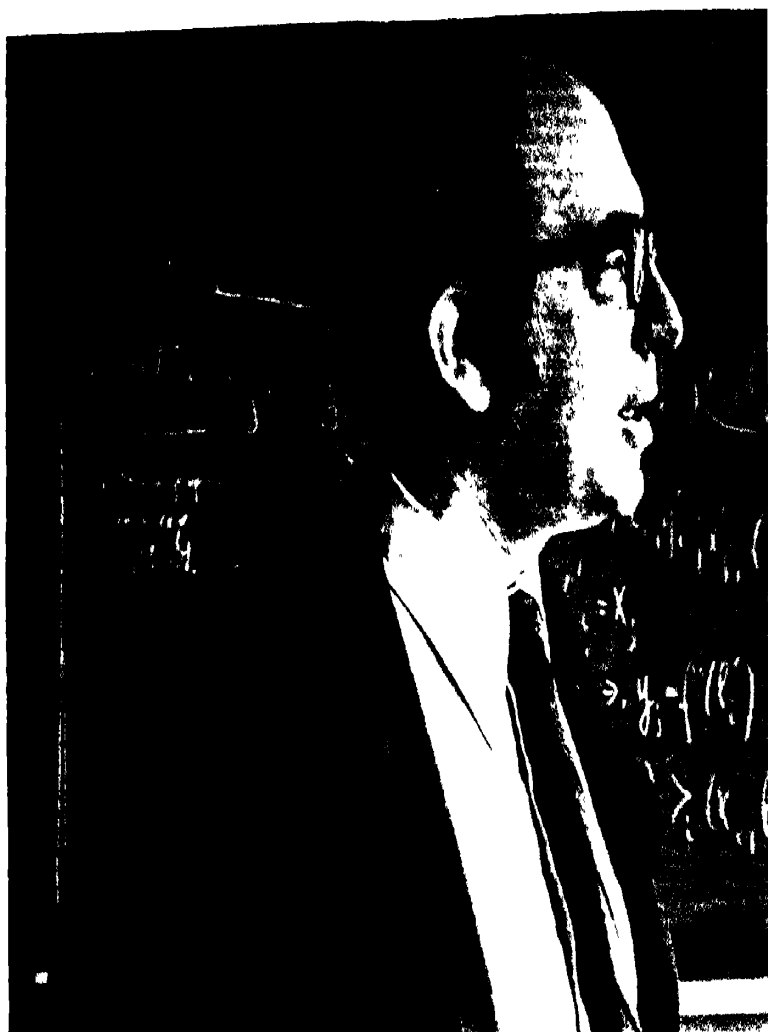
LEONID HURWICZ

Distinguished Fellow

1977

Leonid Hurwicz has brought to economics a rare combination of qualities: breadth of vision along with technical mastery, innovativeness, and a sense of exploration combined with a thorough understanding and appreciation of orthodox theory; and perseverance in he pursuit of unfashionable directions of research together with leadership in laying the oundations for tomorrow's fashions.

His work has encompassed statistical time-series analysis and statistical decision theory - o which he contributed a generalized minimax principle; he participated in the early development of linear programming, and made numerous contributions to non-linear programming and programming in infinite-dimensional spaces; with Kenneth Arrow he developed the general-equilibrium dynamics of the tâtonnement price-adjustment process; with Hirofumi Uzawa he contributed significant conceptual clarification to the elusive ntegrability problem, which has applications in areas as diverse as demand theory and cost-benefit analysis; and he has pioneered a new approach to the study of decentralized esource allocation processes and comparative economic systems in which the adjustment mechanism itself is treated as a decision variable.

An inspiring and witty teacher and colleague, his impatience with pomposity and obfuscation has been matched by his patient encouragement to all who would learn. By insisting hat in deductive reasoning logic admits of no compromise, and by training students to hink clearly and to be suspicious of unclear arguments, he has done much to enhance economics as a science.

THE AMERICAN ECONOMIC REVIEW

June 1978

VOLUME 68, NUMBER 3

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H Borts, Managing Editor, Box Q, Brown University, Providence, R I 02912 Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript A submission fee must accompany each manuscript $15 for members, $30 for nonmembers *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor

• No responsibility for the views expressed by authors in this *REVIEW* is assumed by the editors or the publishers, The American Economic Association

Articles

Shorter Papers

The Technology of Risk and Return

By EDWARD GREENBERG, WILLIAM J. MARSHALL, AND JESS B. YAWITZ*

The behavior of the firm under uncertainty and the valuation of risky assets are basic concerns of contemporary economic and financial theorists. In this study, we present several models which illustrate the interdependence of the micro-economic decisions of the firm and the market's valuation of its capital assets. Such decisions as price and capital investment are shown to affect the risk-return combination presented by the firm to the financial market for valuation. It is assumed that decisions are made in order to maximize the market value of the firm given the risk-return preferences of investors. Thus, the models we present directly link behavior in the product and financial markets.

To represent the financial markets, we make use of the capital asset pricing model (*CAPM*) developed by William Sharpe, John Lintner, and Jan Mossin. This theory provides, as a condition for capital market equilibrium, an explicit model of the valuation of risky assets. Several advantages over the more familiar expected utility approach result from the use of a positive theory of market value. First, it is desirable that the model of firm decision making under uncertainty be consistent with a financial market equilibrium. Second, the goal of the firm appropriate to financial market theory is the maximization of the market value of the firm. In contrast to the expected utility approach, analysis under the *CAPM* does not depend on individuals' preference. Third, the use of an explicit model of asset value enables us to derive explicit and computable comparative static results. The resulting implications for firm behavior appear to be suitable for empirical testing and investigation. The prospects for empirical work are further enhanced by the common dependence of all firms on capital market-determined parameters. Finally, although the point is not developed in this paper, it should be noted that our approach has the potential to link firms' actions with changes in macro-economic activity. To the extent macro changes affect risk preferences, there will be an impact on, for example, investment decisions. In turn, these have macro implications.

We turn to a brief literature review before presenting the basic model and illustrations.

I. Review of Literature

This brief review is specifically concerned with the treatment of risk in the valuation of security prices and its role in investment decisions, beginning with the contribution of Franco Modigliani and Merton Miller.

Modigliani and Miller emphasized the now obvious conclusion that the total value of a firm depends on the nature of its assets and, in the absence of taxes, is independent of financial structure. The financial structure of the firm determines only the allocation of the firm's value between debt and equity investors. To control for the effect of uncertainty on the value of assets, they defined the concept of "equivalent risk classes." Each risk class consists of firms whose returns are perfectly and positively correlated. These firms are considered equally risky in the sense that the uncertain return distribution associated with investment in a firm can be replicated exactly by appropriate investment in any other firm in the same risk class. Being equally risky, all firms in a particular risk class are subject to the same required rate of return on capi-

*Professor of economics, Washington University (St Louis), assistant professor of finance, Texas Christian University, and associate professor of finance and business economics, Washington University. We wish to thank Charles Wilson of the University of Wisconsin and Frederick Warren-Boulton of Washington University for their helpful comments on an earlier version of this paper and Phillip Kott of Brown University for pointing out an error in one of our equations. We take collective responsibility for any remaining errors.

tal and use this rate in making all investment decisions

Unfortunately, the theory does not suggest what determines the absolute, or even relative riskiness, of a particular risk class. Therefore, one is unable to rank classes on the basis of risk and systematically investigate the relationship between risk and the required rate of return on capital

The development of the *CAPM* in the mid-1960's permits, under a particular set of assumptions on investors' preferences and market behavior, a consistent answer to both questions. Rather than presuppose the existence of risk classes, risk is assumed to be a function of the covariance between a firm's cash flows and the market's cash flows. This view of risk is consistent with a normative theory of portfolio selection under the assumption of normally distributed returns. All firms are directly comparable once their returns and covariances are known, and the prices of securities can be determined. Moreover, investment decisions can be based on a risk-adjusted return, consistent with Modigliani and Miller, with the appropriate risk adjustment provided by the theory.[1]

This model is similar to a model of an exchange economy in the sense that the amount of risk and return offered by firms is assumed to be fixed, and the market determines prices for securities in such a way that investors are willing to hold the existing risk and return combinations. Investors' willingness to hold such combinations is reflected in the market price of risk, a parameter of the equation which determines security prices

The use of the *CAPM* as a medium for an investigation of firm behavior is a direct extension of the expected utility approach exemplified by the works of Ira Horowitz, Hayne Leland (1972, 1974), Agnar Sandmo, Raveendra Batra and Aman Ullah, and others. These authors propose that the firm maximizes the expected utility of its owners by its choice over various decision variables. The *CAPM* is based on the assumption that investors are expected utility maximizers, but hold portfolios of investments rather than positions in single firms. While the expected utility approach is more general, it yields results which are dependent on the preferences of an individual and therefore is unable to contend with problems where the firm represents many investors or individuals hold many investments. The *CAPM* requires stronger assumptions about individual preferences, but overcomes these other limitations.

The investment model of Michael Jensen and John Long, discussed below, is a link between the *CAPM* and the present paper.[2] In their study, the firm chooses an optimal rate of investment within the *CAPM* framework Although not explicitly described as such, this choice entails a choice of mean and covariance for the firm, parameters which are ordinarily considered fixed in the *CAPM* Our intention is to extend the idea of choosing mean and covariance to other situations, and to draw some implications for theory and for empirical studies; the general idea is sketched in the next section.

II. Production of Risk and Return

The analysis of capital market behavior which results in the well-known *CAPM* equation,

$$(1)\ E(\tilde{R}_i) = R_f + \frac{E(\tilde{R}_m) - R_f}{\sigma(\tilde{R}_m)} \frac{COV(\tilde{R}_i, \tilde{R}_m)}{\sigma(\tilde{R}_m)}$$

closely parallels the analysis of general equilibrium models of exchange in economics.[3] These models are designed to de-

[1]See the work of Robert Hamada or Mark Rubenstein

[2]The interested reader can gain additional background for our paper from the works of Steiner Ekern and Robert Wilson, Eugene Fama, Robert Merton and Marti Subrahmanyam, Joseph Stiglitz, and our 1975 paper

[3]We employ the conventional *CAPM* notation in equation (1)

$E(\tilde{R}_i)$ = the expected rate of return on security i
R_f = the risk free rate of return
$E(\tilde{R}_m)$ = the expected rate of return on the so-called market portfolio
σ, COV = the standard deviation and covariance operators, respectively

termine market-clearing prices for fixed quantities of products, given consumer preferences. In the *CAPM*, firms offer two financial products, expected return and risk, where risk is proportional to the covariance between the firm's rate of return and the rate of return of the market portfolio. It is assumed that expected return is positively valued and risk is negatively valued by investors. The price of each firm's securities will adjust, given expected value and covariance of cash flows, so that the returns obtained by security holders satisfy equation (1).

Although many interesting and useful results emerge from the analysis of exchange of fixed quantities of risk and return, an equally important set of issues arises in connection with the fact that the firm may vary the risk-return combination it offers to the market by its choice of such variables as price and investment. Thus, in analogy to production theory, the firm maximizes its market value by choosing from among the attainable set of risk-return combinations. That set represents the "technology" for the production of risk and return. We study this production in a partial equilibrium setting.

Since $\tilde{R}_i = \tilde{X}_i/S_i$, where $\tilde{X}_i$ and S_i denote the firm's net income and market value, respectively, we solve equation (1) for S_i as

$$(2) \qquad S_i = \frac{1}{R_f}\,[E(\tilde{X}_i) - \lambda_m COV(\tilde{X}_i, \tilde{X}_m)]$$

$$\text{where } \lambda_m \equiv \frac{E(\tilde{X}_m) - R_f S_m}{\sigma^2(\tilde{X}_m)}$$

S_m is the value, and X_m is the net income, respectively, of all assets comprising the market portfolio; λ_m is the equilibrium rate of exchange of return for risk and is normally referred to as the "market price of risk." Its value reflects the community's aggregate risk-bearing preferences. Although λ_m is considered exogenous to the firm, we consider the effects of changes in its value on several of the firm's decisions. Equation (2) clearly indicates the positive effect of $E(\tilde{X}_i)$ and the negative effect of $COV(\tilde{X}_i, \tilde{X}_m)$ on valuation (assuming $COV > 0$).[4] We now consider the technology producing μ and COV. Figure 1 indicates the general properties that we would like this technology to have. The set of feasible combinations lies under the curve; the curve represents the efficient set where maximum μ is obtained for a given COV. We draw this efficient set as convex to reflect the idea that increases in COV yield successively smaller increase in μ. The location of the intercept is discussed below. As we indicate later, the various models which are developed generate a frontier similar to Figure 1.

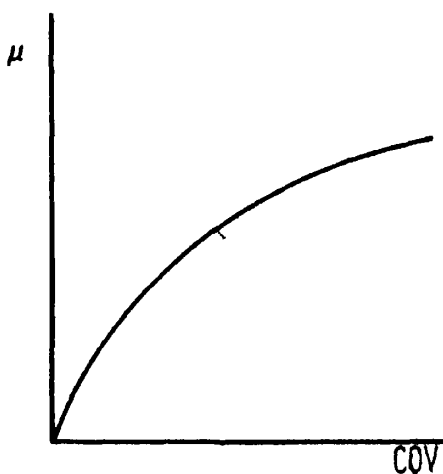



FIGURE 1

If we denote the boundary of the efficient set by

$$(3) \qquad f(\mu, COV) = 0$$

we can maximize equation (2) subject to equation (3).

$$(4) \quad H = \frac{1}{R_f}[\mu - \lambda_m COV] - \gamma f(\mu, COV)$$

$$\frac{\partial H}{\partial \mu} = \frac{1}{R_f} - \gamma f'_{\mu} = 0$$

$$\frac{\partial H}{\partial COV} = -\frac{\lambda_m}{R_f} - \gamma f'_{COV} = 0$$

[4]For the remainder of this section we drop the firm subscript and define

$$\mu = E(\tilde{X}_i), \; COV = COV(\tilde{X}_i, \tilde{X}_m)$$

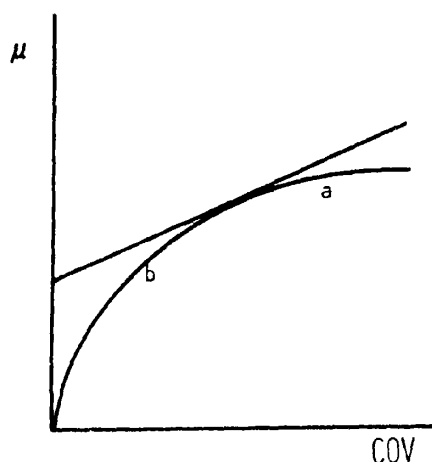



FIGURE 2

which implies

$$(5) \qquad \frac{f'_{COV}}{f'_\mu} = \lambda_m$$

That is, to maximize value, choose the μ-COV combination at which the slope of the efficient set equals the market price of risk. Graphically, we place equations (2) and (3) on Figure 2, moving as far to the northwest as possible. This diagram allows us to examine the effects of changes in the technology and changes in λ_m on the value-maximizing μ-COV bundle. As an example, an increase in the market price of risk will increase the slope of the line and result in a lower μ and a lower COV (a movement from point a to b). We next consider specific examples of the μ-COV technology set.

III. Examples of Risk and Return Production

In this section we illustrate the μ-COV technology by considering several simple models in which a firm attempts to maximize its market value by choosing values for a set of variables under conditions of uncertainty regarding future values of other variables. In each case we examine the sensitivity of the decisions to changes in the market price of risk and other parameters.

A. *Investment under Constant Stochastic Returns to Scale*

We begin with this case because it has already been discussed in the literature, although from a different perspective. Jensen and Long consider a firm investing an amount I, on each dollar of which it earns the random rate of return ρ. The covariance of firm j's cash flow with the market is then

$$(6) \quad COV[X_j + \rho I, X_1 + X_2 + \ldots X_j + \rho I + \ldots X_n] = \sigma_{jm} + I\sigma_{j\rho} + I\sigma_{\rho m} + I^2\sigma_\rho^2$$

where
$$\sigma_{jm} = COV[X_j, X_1 + \ldots + X_n]$$
$$\sigma_{j\rho} = COV[X_j, \rho]$$
$$\sigma_{\rho m} = COV[\rho, X_1 + \ldots + X_n]$$
$$\sigma_\rho^2 = VAR(\rho)$$

Since we wish to avoid questions associated with financing methods, the simplest approach is to assume that the firm acquires its capital equipment by rental, so that the value of the firm must be diminished by this obligation

$$(7) \quad S_j = \frac{1}{R_f}[\mu_j + \bar{\rho}I - \lambda_m(\sigma_{jm} + I\sigma_{j\rho} + I\sigma_{\rho m} + I^2\sigma_\rho^2)] - I$$
$$= \frac{1}{R_f}[\mu_j + (\bar{\rho} - R_f)I - \lambda_m(\sigma_{jm} + I\sigma_{j\rho} + I\sigma_{\rho m} + I^2\sigma_\rho^2)]$$

Jensen and Long proceed by determining the value of I which maximizes S_j:

$$(8) \quad \frac{\partial S_j}{\partial I} = \frac{1}{R_f}[\bar{\rho} - R_f - \lambda_m \cdot (\sigma_{j\rho} + \sigma_{\rho m} + 2I\sigma_\rho^2)] = 0$$

or

$$(9) \quad I^* = \frac{(\bar{\rho} - R_f) - \lambda_m(\sigma_{j\rho} + \sigma_{\rho m})}{2\lambda_m\sigma_\rho^2}$$

Alternatively, since each choice of I determines a different μ-COV combination, we may view the firm as choosing a μ-COV package through its choice of I so as to maximize its value. Let $\mu(I)$ and $COV(I)$

be the respective return and covariance dependent on I. Then, from equation (7)

$$(10)\qquad \mu(I) = \mu + (\bar{p} - R_f)I$$

$$(11)\qquad COV(I) = \sigma_{Im} + I\sigma_{Ip} + I\sigma_{pm} + I^2\sigma_p^2$$

To determine the boundary of the efficient set, we solve equation (10) for I and substitute into (11) to obtain

$$(12)\qquad COV(I) = \sigma_{Im} + \left(\frac{\mu(I) - \mu}{\bar{p} - R_f}\right) \cdot (\sigma_{Ip} + \sigma_{pm}) + \sigma_p^2\left(\frac{\mu(I) - \mu}{\bar{p} - R_f}\right)^2$$

$$= \left[\sigma_{Im} - \frac{\mu(\sigma_{Ip} + \sigma_{pm})}{\bar{p} - R_f} + \frac{\mu^2\sigma_p^2}{(\bar{p} - R_f)^2}\right] + \frac{\mu(I)}{\bar{p} - R_f}\left[\sigma_{Ip} + \sigma_{pm} - \frac{2\mu\sigma_p^2}{\bar{p} - R_f}\right] + \mu^2(I)\frac{\sigma_p^2}{(\bar{p} - R_f)^2}$$

$COV(I)$ is thus quadratic in $\mu(I)$ and is therefore consistent with Figure 1; the point at which it crosses the horizontal axis depends on the values of the parameters.

From (12), using (5) we can calculate

$$(13)\qquad -\frac{f_{COV(I)}}{f_{\mu(I)}} = 1 \div \left[\frac{\sigma_{Ip} + \sigma_{pm}}{\bar{p} - R_f} + \frac{2\sigma_p^2(\mu(I) - \mu)}{(\bar{p} - R_f)^2}\right] = \lambda_m$$

It is easy to verify that (13) implies the same value for I as equation (9). Thus the optimal amount of investment in this model may be interpreted as the choice of a value-maximizing μ-COV package, given market parameters. It is clear from (9) that an increase in λ_m, σ_{pm}, R_f, σ_{Ip} or σ_p^2 will decrease investment, while an increase in $\bar{p}$ will increase it. This model might be exploited for the explanation of investment decisions by explicitly bringing in the effects of risk.

B. *Proportional Expansion in an Industry*

We next consider the example given by Mossin of the effect on valuation of proportional expansion of a firm's cash flows. Again this will be interpreted as an optimal μ-COV choice by the firm.

With initial expected value μ, expansion by 100γ% is assumed to lead to a new expected value $\mu(\gamma)$, where

$$(14)\qquad \mu(\gamma) = (1 + \gamma)\mu$$

As Mossin shows, the new covariance, $COV(\gamma)$, is equal to

$$(15)\qquad (1 + \gamma)COV + \gamma(1 + \gamma)\sigma^2$$

where COV is the covariance of the firm's cash flows with the market before expansion, and σ^2 is the variance of the firm's own cash flows. It can be shown that the optimal γ is given by

$$(16)\qquad \gamma^* = \frac{\mu - \lambda_m(COV + \sigma^2)}{2\lambda_m\sigma^2}$$

Note that γ^* may be negative, indicating a decrease in the firm's scale and that

$$\partial\gamma^*/\partial\lambda < 0,$$
$$\partial\gamma^*/\partial\mu > 0,$$

and

$$\partial\gamma^*/\partial\sigma^2 \gtrless 0 \text{ as } COV \gtrless \mu/\lambda_m$$

As above, the relation between $\mu(\gamma)$ and $COV(\gamma)$ is quadratic. Figure 3 indicates the general shape for two cases according to whether $COV > \sigma^2$ and $COV < \sigma^2$, respectively. The frontier for the latter is derived by permitting γ to increase from -1. It implies that a firm capable of "producing" a negative covariance will not do so for sufficiently low values of λ_m. This may explain why negative covariance is rarely observed. Empirical work which examines the behavior of firms in severe depressions, during which λ is likely to be high, may reveal instances of negative covariance.

C. *Fixed and Variable Costs: Competitive Firm*

The next three models are similar to those presented by Horowitz (chs. 12 and 13) who utilized the expected utility maximization hypothesis as well as assumptions about the risk-aversion characteristics of

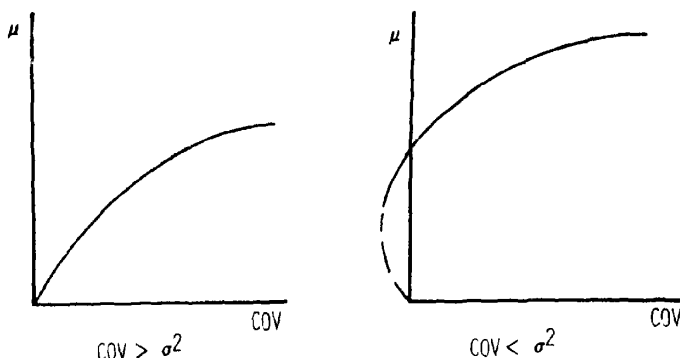



FIGURE 3

the decision maker.

In this example we assume that the firm faces a random price which is not affected by its output, and can adjust its output when price is known. Capital stock (K), however, must be chosen before price is known. We assume a Cobb-Douglas production function in capital and labor (L), so that output (Q) is given by

$$(17) \qquad Q = L^{\alpha}K^{1-\alpha}$$

For a given capital stock (K_o), the minimum labor required to produce Q is

$$(18) \qquad L(Q;K_o) = Q^{1/\alpha}K_o^{(\alpha-1)/\alpha}$$

Total cost, assuming a wage rate of w and unit cost of capital of r, is given by

$$(19) \quad C(Q;K_o) = wQ^{1/\alpha}K_o^{(\alpha-1)/\alpha} + rK_o$$

Since the average variable cost equals zero for $Q = 0$ the firm will produce at any positive price. Profits are given by

$$(20) \quad \pi = PQ - wK_o^{(\alpha-1)/\alpha}Q^{1/\alpha} - rK_o$$

and maximum profits for a given P are obtained for

$$(21) \qquad Q = K_o\left(\frac{\alpha P}{w}\right)^{\alpha/(1-\alpha)}$$

Maximum profits, given P, are

$$(22) \qquad \pi_o = K_o(C_1 P^{1/(1-\alpha)} - r)$$

where $C_1 = (1-\alpha)\left(\frac{\alpha}{w}\right)^{\alpha/(1-\alpha)}$

The expected value of maximum profits[5] is

$$(23) \quad E(\pi_o) = K_o[C_1 E(P^{1/(1-\alpha)}) - r] = K_o C_2$$

and it is assumed that $C_2 > 0$.

To determine the covariance of this firm's profits with the market, we define W' as the income produced by the economy other than this firm, and W as the total income.[6] That is,

$$(24) \qquad W = W' + K_o \cdot [C_1 P^{1/(1-\alpha)} - r]$$

Then the covariance between π and W is

(25)

$$COV(\pi, W) = E[\pi - E(\pi)][W - E(W)]$$
$$= E\{[K_o C_1(P^{\delta} - E(P^{\delta})][W' - E(W') + K_o C_1 P^{\delta} - E(P^{\delta})]\}$$
$$= K_o^2 C_1^2 VAR(P^{\delta}) + K_o C_1 COV(P^{\delta}, W')$$

[5]Note that we are assuming that for a given price the firm will maximize profits. This may be inconsistent with the goal of maximizing value since profits also enter into covariance. This problem appears to require a control-theoretic approach, and will be the subject of future research.

[6]We assume that W' affects P, but that there is no effect on W' from the firm's choice of K. In effect, we are assuming that $P = \beta + \rho W' + u$, where W' and u are random variables which are mutually independent and independent of K.

where $\delta = 1/(1-\alpha)$

We derive the relationship between $E(\pi)$ and $COV(\pi, W)$ by solving each for K and equating the two terms. In the quadratic expression for $COV(\pi, W)$ we take the larger of the two roots, since for a given covariance, the larger root yields the larger expected value. Thus

$$(26) \quad K = [-COV(P^\delta, W') + (COV^2(P^\delta, W') + 4\, COV(\pi, W)VAR(P^\delta))^{1/2}] \div 2C_1 VAR(P^\delta)$$

Equating the above to $K = E(\pi)C_2$, we obtain the following expression for the mean-covariance frontier:

$$(27) \quad E(\pi) = [-C_2 COV(P^\delta, W') + C_2 \cdot (COV^2(P^\delta, W') + 4\, COV(\pi, W)\, VAR(P^\delta))^{1/2}] \div 2C_1\, VAR(P^\delta)$$

It is easy to verify that

$$\frac{\partial E(\pi)}{\partial COV(\pi, W)} > 0 \text{ and } \frac{\partial^2 E(\pi)}{\partial COV^2(\pi, W)} < 0$$

so that the μ-COV frontier has the correct slope and curvature.

The optimal value for K is most readily obtained directly from the expression for the value of the firm:

$$(28) \quad R_f S = K_o C_2 - \lambda_m [K_o^2 C_1^2\, VAR(P^\delta) + K_o C_1 COV(P^\delta, W')]$$

Letting $dR_f S/dK_o = 0$, we obtain

$$(29) \quad K = \frac{C_2 - \lambda_m C_1 COV(P^\delta, W')}{2\lambda_m C_1^2 VAR(P^\delta)}$$

Evidently, $\partial K_o/\partial\lambda_m < 0$ and

$$\frac{\partial K_o}{\partial COV(P^\delta, W')} < 0$$

Thus, as the market price of risk increases, the firm will operate with a smaller capital stock, that is, use more labor to produce the same output. This will increase the amount of variable costs relative to fixed costs. The same effect arises from an increase in $VAR(P^\delta)$ and $COV(P^\delta, W')$, which seems reasonable: increases in riskiness or distaste for risk lead to an attempt to smooth profits by reducing fixed costs.

D. *Monopolist Firm I: Output as Decision Variable*

In this case, a monopolist firm chooses output before price is known. We again assume a Cobb-Douglas production function; in addition, price is given by

$$(30) \quad P = Q^\beta Y^\gamma, \qquad -1 < \beta < 0$$

where the random variable Y denotes the measure of aggregate market performance employed in the $CAPM$, $-\beta$ is the inverse of the firm's demand elasticity, and γ measures the response of demand to Y.

For a given value of output, $\overline{Q}$, minimum costs are given by

$$(31) \quad C(\overline{Q}) = \frac{r}{1-\alpha}\left(\frac{(1-\alpha)w}{\alpha r}\right)^\alpha \overline{Q} = C\overline{Q}$$

Then profits are given by

$$(32) \quad \pi = P\overline{Q} - C(\overline{Q}) = \overline{Q}^{\beta+1} Y^\gamma - C\overline{Q}$$

The expected value of profits is needed for the valuation equation. It is given by

$$(33) \quad E(\pi) = \overline{Q}^{\beta+1} E(Y^\gamma) - C\overline{Q}$$

Moreover,

$$(34) \quad COV(\pi, Y) = COV(P\overline{Q}, Y) = \int \overline{Q}^{\beta+1} Y^\gamma (Y - E(Y)) \cdot f(Y)dY = \overline{Q}^{\beta+1} COV(Y^\gamma, Y)$$

Accordingly, the firm's value is given by

$$(35) \quad V = \frac{1}{R_f}[\overline{Q}^{\beta+1}(E(Y^\gamma) - \lambda_m COV(Y^\gamma, Y)) - C\overline{Q}]$$

We choose Q^* to maximize V as

$$(36) \quad Q^* = C^{1/\beta} \div ((\beta+1)[E(Y^\gamma) - \lambda_m COV(Y^\gamma, Y)])^{1/\beta}$$

Since the second derivative of V with respect to Q yields a negative value (on the

assumption that $E(Y^\gamma) - \lambda_m COV(Y^\gamma, Y) > 0$, a necessary condition for S to be positive), the frontier must have the desired shape

If $COV(Y^\gamma, Y) > 0$, an increase in λ_m will reduce Q. Although an increase in λ_m reduces V, the ability to adjust Q and thereby $E(\pi)$ and $COV(\pi, Y)$ results in a smaller decrease in value than that which would have occurred if Q had held at its initial value. This is the main distinction between our "production economy" approach and the conventional analysis conducted within an exchange economy framework. Note that this approach links changes in underlying economic parameters to changes in output and investment decisions through the security market valuation process.

Returning to equation (36), we see that the less elastic is demand, the smaller will be Q (if $Q^* > 1$) and the higher will be price An increase in C (as a result of an increase in w or r) will decrease Q and the use of inputs. The effect of γ on Q is ambiguous. Specifically, $\partial Q/\partial\gamma \gtreqless 0$ depending on whether

$$\frac{\partial E(Y^\gamma)}{\partial\gamma} \bigg/ \frac{\partial COV(Y^\gamma, Y)}{\partial\gamma} \gtrless \lambda_m$$

E. *Monopolist Firm II*
Price as Decision Variable

In this case, we consider a firm which sets price and chooses its capital stock before the quantity demanded, a random variable, is known. In a somewhat different setting, Robert Meyer has considered a similar case where the monopolistic firm chooses the prices to be charged to various customers in order to maximize its equity value under the *CAPM* In our example, when the quantity demanded becomes known, the firm hires sufficient labor to produce it. A public utility which sets price and is required to meet demand would be an example of the type of firm covered by this model. We assume the same production and demand functions as in the previous section. The main results are given in the equations for P and K:

$$(37) \quad P = \left(\frac{w}{\alpha}\right)^{\alpha}\left(\frac{r}{1-\alpha}\right)^{1-\alpha}\frac{1}{1+\beta}\frac{C_1^{\alpha}}{C_2}$$

$$(38) \quad K = \left(\frac{w(1-\alpha)}{r\alpha}\right)^{\alpha(1+1/\beta)} \cdot \left(\frac{r}{(1+\beta)(1-\alpha)}\right)^{1/\beta} C_1^{\alpha(1+1/\beta)} \cdot C_2^{-1/\beta}$$

where

$$C_1 = E(Y^{-\gamma/\alpha\beta}) - \lambda_m COV(Y^{-\gamma/\alpha\beta}, Y)$$
$$C_2 = E(Y^{-\gamma/\beta}) - \lambda_m COV(Y^{-\gamma/\beta}, Y)$$

Because of the complex way in which the parameters (α, β, γ, and moments of the distribution of Y) enter equations (37) and (38), it is difficult to make clear-cut statements regarding their impacts on the variables of interest. This is in contrast to the previous section in which inputs and outputs were chosen before price became known In that case, specific implications could be drawn. However, when K alone is chosen prior to receiving price information, even qualitative results depend on parameter variables. It is possible to show that $E(K/Q)$ varies inversely with both λ_m and $COV(Y^{-\gamma/\alpha\beta}, Y)$ if the latter is positive.

F. *Diversification*

We conclude with a discussion of proportional expansion by considering a firm which operates in two industries, and which is contemplating expansion in both at rates γ_1 and γ_2, respectively As we shall see, the presence of negative covariance between the two industries provides incentive for diversification at the firm level. The new expected value of the net cash flows are

$$(39) \quad \mu' = (1 + \gamma_1)\mu_1 + (1 + \gamma_2)\mu_2$$

and the new covariance is

$$COV' = (1 + \gamma_1)\sigma_{1m} + (1 + \gamma_2)\sigma_{2m} + \gamma_1(1 + \gamma_1)\sigma_{11} + \gamma_2(1 + \gamma_2)\sigma_{22} + (\gamma_1 + \gamma_2 + 2\gamma_1\gamma_2)\sigma_{12}$$

where

μ_i = expected value of the cash flows in industry i

σ_{im} = covariance between industry i and the market

σ_{ij} = covariance between industry i and j

$(i, j = 1,2)$

The presence of σ_{12} in the expression for COV' is interesting. If the value-maximizing rule indicates positive values for γ_1 and γ_2, and if $COV < 0$, the covariance for simultaneous expansion in the two industries will be lower than that for each taken separately on the assumption that an individual firm takes the sizes of other firms as given (a Cournot-type assumption regarding expansion decisions). This creates an externality in the sense that a firm operating in industry 1 confers a benefit on firms in industry 2 when it expands. The uncoordinated activities of the two firms are not likely to achieve the rate of expansion which maximizes the sum of the market value of the two firms.

Moreover, within the framework of the *CAPM*, stockholders who recognize the negative covariance cannot induce the firms to achieve optimal size by affecting securities prices.[7] This is so in that model because equilibrium prices depend only upon the risk and return combination offered by the firm, and the firm would regard a departure from the equilibrium level as a temporary aberration, not as a signal to expand. To take advantage of the benefits from negative covariance as described here, it is necessary that a single firm control the two decisions. Of course, it may be possible to conceive of arrangements in which one firm bribes another to produce at a particular output, but this is difficult because of antitrust laws. Another interesting feature of this example is that it may explain the lack of examples of observed negative covariance. If firms diversify to take advantage of it, a considerable amount of negative covariance will be reflected *within* firms, and not show up when correlations between firms are examined. Finally, note that this type of diversification can reduce the amount of risk in the economy, and therefore has beneficial aspects. Of course, it may also have anticompetitive aspects.

We may obtain the optimal values of γ_i by differentiating the expression for the value of the firm. The first-order condition for γ_1 is

$$(40)\quad R_f\frac{\partial S}{\partial \gamma_1} = \mu_1 - \lambda_m[\sigma_{1m} + (1 + 2\gamma_1)\sigma_{11} + (1 + 2\gamma_2)\sigma_{12}] = 0$$

which yields

$$(41)\quad \gamma_1 = \frac{1}{2}\left\{\left|\sum\right|^{-1}\left[\sigma_{22}\left(\frac{\mu_1}{\lambda_m} - \sigma_{1m}\right) - \sigma_{12}\left(\frac{\mu_2}{\lambda_m} - \sigma_{2m}\right)\right] - 1\right\}$$

where $\sum = \begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}$

The sensitivity of the optimal γ's to the system parameters may be examined by total differentiation of the first-order conditions shown in equation (42). We can then derive the partial derivatives shown on the next page.

[7] We are indebted to Warren-Boulton for helpful discussion on this point.

$$(42)\quad \begin{pmatrix} d\gamma_1 \\ d\gamma_2 \end{pmatrix} = \frac{1}{2}\sum^{-1}\begin{bmatrix} \frac{1}{\lambda_m}\left(d\mu_1 - \frac{\mu_1}{\lambda_m}\,d\lambda_m\right) - d\sigma_{1m} - (1 + 2\gamma_1)d\sigma_{11} - (1 + 2\gamma_2)d\sigma_{12} \\ \frac{1}{\lambda_m}\left(d\mu_2 - \frac{\mu_2}{\lambda_m}\,d\lambda_m\right) - d\sigma_{2m} - (1 + 2\gamma_2)d\sigma_{22} - (1 + 2\gamma_2)d\sigma_{12} \end{bmatrix}$$

$$\frac{\partial\gamma_1}{\partial\mu_1} = \frac{\sigma_{22}}{2|\Sigma|\lambda_m} > 0, \frac{\partial\gamma_1}{\partial\mu_2} = \frac{-\sigma_{12}}{2|\Sigma|\lambda_m},$$

$$\frac{\partial\gamma_1}{\partial\lambda_m} = \frac{\sigma_{12}\mu_2 - \sigma_{22}\mu_1}{2|\Sigma|\lambda_m^2}$$

$$\frac{\partial\gamma_1}{\partial\sigma_{11}} = \frac{-\sigma_{22}(1+2\gamma_1)}{|\Sigma|} < 0,$$

$$\frac{\partial\gamma_1}{\partial\sigma_{22}} = \frac{\sigma_{12}(1+2\gamma_2)}{|\Sigma|},$$

$$\frac{\partial\gamma_1}{\partial\sigma_{12}} = \frac{\sigma_{12}(1+2\gamma_2) - \sigma_{11}(1+2\gamma_1)}{|\Sigma|}$$

If $\sigma_{12} > 0$, we have $\frac{\partial\gamma_1}{\partial\sigma_{22}} > 0, \frac{\partial\gamma_1}{\partial\mu_2} < 0$

The signs of the other derivatives are ambiguous:

$$\frac{\partial\gamma_1}{\partial\lambda_m} \gtreqless 0 \text{ as } \sigma_{12}\mu_2 - \sigma_{22}\mu_1 \gtreqless 0$$

and $\frac{\partial\gamma_1}{\partial\sigma_{12}} \gtreqless 0$

$$\text{as } \sigma_{12}(1+2\gamma_2) - \sigma_{11}(1+2\gamma_1) \gtreqless 0$$

If $\sigma_{12} < 0$,

$$\frac{\partial\gamma_1}{\partial\lambda_m} < 0, \frac{\partial\gamma_1}{\partial\sigma_{12}} < 0, \frac{\partial\gamma_1}{\partial\sigma_{22}} < 0, \frac{\partial\gamma_1}{\partial\mu_2} > 0$$

Whether or not the firm grows as a result of a parameter change may be determined by examining the partial derivative of μ' with respect to the particular parameter. A positive value indicates an increase in expected value, hence, an increase in the scale of the firm. Consider the effect of a change in the market price of risk:

$$(43)\quad \frac{\partial\mu'}{\partial\lambda_m} = \mu_1\frac{\partial\gamma_1}{\partial\lambda_m} + \mu_2\frac{\partial\gamma_2}{\partial\lambda_m}$$

$$= \mu_1\left(\frac{\sigma_{12}\mu_2 - \sigma_{22}\mu_1}{2|\Sigma|\lambda_m^2}\right) + \mu_2\left(\frac{\sigma_{12}\mu_1 - \sigma_{11}\mu_2}{2|\Sigma|\lambda_m^2}\right)$$

$$= -\frac{1}{2|\Sigma|\lambda_m^2}(\sigma_{22}\mu_1^2 + \sigma_{11}\mu_2^2 - 2\mu_1\mu_2\sigma_{12})$$

$$< 0$$

Thus, an increase in the market price of risk will always reduce the scale of the firm.

IV. Implications and Extensions

In this paper we have attempted to link the capital asset pricing model, which has been extensively studied in the finance litterature, with the type of pricing, output, and investment models extensively studied in the economics literature. We considered a number of models in which a firm attempting to maximize its value as determined by the *CAPM* would be led to particular choices for those decision variables. Our results demonstrate the interaction of the firm's operating environment and the risk-return preferences of the financial market in the determination of the firm's value-maximizing behavior. The analysis produced solutions for optimal values of decision variables which depend on potentially measurable firm and financial market characteristics. Therefore, there would seem to exist an opportunity to use this approach to generate empirically testable hypotheses.

In addition, it may be of interest to enlarge the set of decision variables to include some parameters which were previously considered exogenous to the firm. As an example, value-maximizing behavior may dictate that a firm attempt to change its product's price and income elasticities through advertising or quality changes. A firm may undertake an advertising policy designed to increase the degree to which consumers view its product to be a necessity. If successful, such a policy would, *ceteris paribus*, reduce the product's income elasticity and bring about a decrease in the firm's risk. Similarly, models of induced technological change might be utilized to investigate the best strategy to be pursued for improvement in technology for a value-maximizing firm.

Finally, we have noted that our approach has the potential to link firm's actions and changes in macro-economic activity. To the extent that the macro functioning of the economy affects risk preferences, there will be an impact on firms' optimal microeconomic decisions. In turn, such induced

changes in firms' behavior have macro implications.

REFERENCES

R. Batra and A. Ullah, "Competitive Firm and the Theory of Input Demand Under Price Uncertainty," *J. Polit. Econ.*, May/June 1974, *82*, 537-48.

S. Ekern and R. Wilson, "On the Theory of the Firm in an Economy with Incomplete Markets," *Bell J. Econ.*, Spring 1974, *5*, 171-80.

E. Fama, "Perfect Competition and Optimal Production Decisions under Uncertainty," *Bell J. Econ.*, Autumn 1972, *3*, 509-30.

R. Hamada, "Portfolio Analysis, Market Equilibrium and Corporation Finance," *J. Finance*, Mar. 1968, *23*, 13-31.

Ira Horowitz, *Decision Making and the Theory of the Firm*, New York 1970.

M. Jensen and J. Long, "Corporate Investment under Uncertainty and Pareto Optimality in the Capital Markets," *Bell J. Econ.*, Spring 1972, *3*, 151-74.

H. Leland, "Theory of the Firm Facing Random Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278-91.

———, "Production Theory and the Stock Market," *Bell J. Econ.*, Spring 1974, *5*, 125-44.

J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *47*, 13-37.

W. Marshall, J. Yawitz, and E. Greenberg, "On the Comparative Statics of Asset Price Adjustments," work. paper, Washington Univ. 1975.

R. Merton and M. Subrahmanyam, "The Optimality of a Competitive Stock Market," *Bell J. Econ.*, Spring 1974, *5*, 145-70.

R. Meyer, "Risk-Efficient Monopoly Pricing for the Multiproduct Firm," *Quart. J. Econ.*, Aug. 1976, *90*, 461-74.

F. Modigliani and M. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261-97.

Jan Mossin, *Theory of Financial Markets*, Englewood Cliffs 1973.

M. Rubinstein, "A Mean-Variance Synthesis of Corporate Financial Theory," *J. Finance*, Mar. 1973, *28*, 167-82.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65-73.

W. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425-42.

J. Stiglitz, "On the Optimality of the Stock Market Allocation of Investment," *Quart. J. Econ.*, Feb. 1972, *86*, 25-60.

Self-Financing of an R&D Project

By MORTON I. KAMIEN AND NANCY L. SCHWARTZ*

Among the characteristics commonly associated with industrial research and development, one of the most prominent is the virtual necessity for it to be financed internally from a firm's current profits and accumulated funds. This feature of *R&D* underlies the view that a firm must possess some monopoly power, along with the associated monopoly profits, for it to carry on *R&D*. Two reasons for this self-financing are frequently offered. First, external financing may be difficult to obtain without substantial related tangible collateral to be claimed by the lender if the project fails. An *R&D* project that fails generally leaves behind few tangible assets of value. Second, the firm might be reluctant to reveal detailed information about the project that would make it attractive to outside lenders, fearing its disclosure to potential rivals. These observations are supported by anecdotal evidence and case studies documenting the difficulties of obtaining adequate external financing for *R&D* and innovation. Nevertheless, regression tests generally lend scant support to the hypothesis that corporate financial liquidity is conducive to innovation (see F. M. Scherer, pp. 363-64, or the authors, 1975, pp. 24-26).

Motivated by the conflicting empirical evidence on the role of financial liquidity in innovation, we seek further insight through a theoretical exploration of optimal *R&D* spending by an expected profit-maximizing firm that must finance its entire effort internally.[1] As a result, a possible resolution of the apparent conflict between the case study and regression analyses is suggested. In particular, we show that an established firm earning profits on its current product can easily finance development of a new product that will, say, double expected profits. Thus for established firms doing "routine" *R&D* to strengthen product lines, the financial constraint should not be binding and regression studies should find no relationship between liquidity and innovative activity. On the other hand, very large innovations or innovations of a new or marginal firm would be constrained by cash availability; it is such innovations that are the subject of anecdotes and case studies.

We envision a firm contemplating new product development to replace its current product and enhance profits. Alternatively, it recognizes that its product might sometime be displaced by some rival product and hopes to resist this possible loss of profits by developing and marketing a new product of its own. The Industrial Research Institute refers to such *R&D* projects as in "support of existing business" (see A. E. Brown). Thus the purpose of *R&D* may be either "offensive" or "defensive."

The firm must determine whether new product development is worthwhile and, if so, an introduction date and spending plan (employing only internally generated funds) that will maximize its discounted expected profits. Since competing development plans are unknown and development of a new product already preempted by another is assumed worthless, profits can be known only probabilistically.

We are able to characterize the optimal development plan by employing methods of optimal control theory. We examine the influences on the optimal development plan of innovational rivalry, current profits, initial cash balances, profits associated with the new product, and the effort required to successfully develop the new product. Four

*Graduate School of Management, Northwestern University. The helpful suggestions of George Borts and a referee are gratefully acknowledged.

[1]Although a firm may be able to obtain external financing for an *R&D* project if it has a good record in such ventures, we restrict attention here to situations of internal funding only. Analysis of a similar problem with external financing readily available was conducted by the authors (1972).

propositions describe some of the major conclusions.

In the next section the details of the model are presented. This is followed by a description of the general solution. The development plan when the cash constraint is inactive and active, respectively, is described in the succeeding two sections. The results are reviewed in the final section.

I. The Model

A. Glossary

Nonnegative parameters:

- π = current profit rate
- P = gross capitalized value of innovation
- r = discount rate = earnings rate on cash balances
- h = intensity of innovational rivalry; conditional probability density of rival entry at any t
- a = degree of diminishing returns to faster $R\&D$ spending
- A = effective development effort required for innovation
- R_0 = initial cash balance

Decision variables:

- $y(t)$ = $R\&D$ spending rate at t
- $z(t)$ = cumulative effective $R\&D$ effort by t
- $R(t)$ = cash balance at t
- T = planned new product introduction date

Multipliers in solution

- λ = implicit marginal cost of effective $R\&D$ effort
- k = implicit marginal cost of self-financing constraint

B

The firm earns profits at the constant rate $\pi \geq 0$ per unit time from the sales of its current product. These profits continue until the product is displaced by the firm's new product or by appearance of a rival substitute product. The class of potential rivals is large and diffuse, possibly including some firms that are currently in the same line of business, firms in other businesses, and newcomers. The firm knows neither the composition of this group of potential rivals, nor precisely when a rival product will be introduced, nor by whom. Its beliefs regarding the introduction date of a rival product are summarized by a probability distribution $F(t)$, where $F(t)$ is the probability that a rival product will appear by time t. In particular we assume the exponential form $F(t) = 1 - e^{-ht}$. The conditional probability of rival product introduction at any time t, given that it has not yet appeared, is $F'/(1 - F) = h$, a constant, and the expected introduction date of the rival product is $1/h$. Thus the parameter h, often called the hazard rate, reflects the intensity of innovational rivalry perceived by the firm in the sense that a higher value of h is associated with the expectation of more imminent introduction of a rival product.

If new product development is not undertaken, the firm receives profits on its current product until a new rival product appears. Then the expected present value of profits from the firm's existing product would be

$$(1) \qquad \int_0^\infty e^{-rt}\pi(1 - F(t))\,dt = \int_0^\infty e^{-(r+h)t}\pi\,dt = \pi/(r + h)$$

where $r > 0$ denotes the earnings rate on cash balances. The profit stream available from product innovation is assumed to have value P when discounted to the moment of introduction. The capital value P may occur in a variety of ways. For instance, a profit stream $p(t)$ may be anticipated and so

$$P = \int_T^\infty e^{-r(t-T)}p(t)\,dt$$

where T is the moment of introduction of the new good. Specifically, the firm might expect to receive a constant flow P_0 for the life L of a patent and nothing thereafter; in

this case

$$P = \int_{T}^{T+l} e^{-r(t-T)} P_0 dt$$

Or the firm might receive P_0 while it is the sole producer of the new good and P_1 once rivals have appeared, then

$$P = \int_{T}^{\infty} e^{-r(t-T)}[P_0(1 - F(t)) + P_1(F(t) - F(T))]dt/[1 - F(T)]$$

This expectation reflects the rewards available and the probabilities with which they will be collected (P_0 is received at t if no rival has entered by t, P_1 is collected at t if a rival has appeared after T but before t). It is not necessary to specify the form or duration of the profit stream from the innovation, its expected capital equivalent P upon innovation is sufficient. Since $1 - F(T)$ is the probability rival preemption has not occurred by time T, the expected reward at T is $P(1 - F(T)) = e^{-hT}P$

For new product development to be attractive, the reward from innovating must exceed the expected profit from failing to do so. We assume therefore that

$$(2) \qquad P > \pi/(r + h)$$

Notice that this is affected by the intensity of innovational rivalry A project that might not be considered at all in the absence of rivalry ($P < \pi/r$) may nonetheless be considered or even undertaken as a defensive measure if rival entry is thought likely

Development of the new product requires accumulation of effective development effort, achieved by efficacious expenditure of money through time While no $R\&D$ project is truly certain in detail, no substantive technical uncertainty is assumed in this problem. We assume that the approximate amount of effort required can be estimated reasonably closely. This assumption is appropriate for a wide class of product development projects (and helps keep the analysis tractable without serious loss of generality).[2] Letting $z(t)$ denote the effective development effort accumulated by time t and $y(t)$ be the spending rate at t, we assume that

$$(3) \qquad z'(t) = y^a(t), \quad z(0) = 0, \quad z(T) = A \text{ where } 0 < a < 1$$

Effective effort is initially zero, accumulates as a concave monotone function of development spending (reflecting diminishing returns to faster spending found in empirical studies of Scherer and Edwin Mansfield et al.), and must achieve a known level A for successful development by any time T.

Let $R(t)$ denote the firm's cash balance at time t. It is augmented by interest earnings at rate r on the principal R and from profit π on the current product; it is diminished by expenditures $y(t)$ on $R\&D$. Hence the changing state of R is described by the differential equation

$$(4) \qquad R'(t) = rR(t) + \pi - y(t), \quad R(0) = R_0 \geq 0$$

Combining assumptions about rewards from the current and new products, the development function, and cash balances, we can state the firm's expected profit-maximization problem. A planned introduction date $T^* > 0$ and development expenditure plan $y^*(t) \geq 0$, $0 \leq t \leq T^*$ are to be chosen to

$$(5) \quad \text{Max} \int_0^T e^{-rt}(\pi - y(t))(1 - F(t))dt + e^{-rT}P(1 - F(T)) = \int_0^T e^{-(r+h)t}(\pi - y(t))dt + e^{-(r+h)T}P$$

subject to (3), (4), and

$$(6) \qquad R(t) \geq 0, 0 \leq t \leq T$$

The objective functional (5) involves the firm's profit from its existing product and expenditures on new product development so long as no new rival product has appeared, as well as the reward from the new product provided no rival product appears before T. If a rival product does appear, the

[2]See the authors (1974) for a theoretical study of a case of important technical uncertainty in $R\&D$.

firm receives and spends nothing thereafter.[3] Although nonnegative cash balances are required at all times, it suffices to require only

$$(7) \qquad R(\mathrm{T}) \geq 0$$

The proof rests on a result to be developed later in Proposition 1 that the solution to (3)–(5), and (7) involves $y'(t) > 0$.[4] A solution in which $T^* \to \infty$ indicates that new product development is not worthwhile.

II. The General Solution

To solve the problem posed in (3)–(5) and (7), we associate multipliers λ and γ with constraints (3) and (4), respectively, and form the Hamiltonian

$$H = e^{-(r+h)t}(\pi - y) + \lambda y^a + \gamma(rR + \pi - y)$$

An optimal solution in which $T^* < \infty$ must satisfy (3), (4), (7), and (8)–(12):

$$(8) \qquad \partial H/\partial y = -e^{-(r+h)t} + a\lambda y^{a-1} - \gamma = 0$$

$$(9) \qquad \lambda' = -\partial H/\partial z = 0$$

so that λ is constant

$$(10) \qquad \gamma' = -\partial H/\partial R = -r\gamma$$

so that $\gamma(t) = ke^{-rt}$

where

$$(11) \qquad k \geq 0, \qquad kR(\mathrm{T}) = 0$$

$$(12) \qquad H(\mathrm{T}) = (r + h)e^{-(r+h)\mathrm{T}}P$$

Conditions (11) follow from the required nonnegativity of $R(\mathrm{T})$; see Kenneth Arrow and Mordecai Kurz. Expression (12) is the transversality condition. Since (3) (5) is concave in y and R, these necessary conditions are also sufficient for optimality.

The solution to the necessary conditions can be summarized in the following four equations. Substituting from (10) into (8) gives

$$(13) \qquad y(t) = \{a\lambda e^{(r+h)t}/(1 + ke^{ht})\}^{1/(1-a)}$$

Substituting from (10), (11), and (13) into (12) gives

$$(14) \qquad (1 + ke^{h\mathrm{T}})(\pi + y(\mathrm{T})(1 - a)/a) = (r + h)P$$

In addition (3), (4) and (7) must be satisfied:

$$(15) \qquad \int_0^{\mathrm{T}} y^a(t)dt = A$$

$$(16) \qquad k[R_0 + (1 - e^{r\mathrm{T}})\pi/r - \int_0^T e^{-rt}y(t)dt] = 0$$

where (11) has been employed in deriving (16). With $y(t)$ specified in (13), the three equations (14)–(16) jointly determine the three nonnegative constants T, λ, and k. Nonnegativity of λ is implied by (13).

Before pursuing detailed analysis of the solution, we can immediately obtain a qualitative property of the optimal spending path from (13).

PROPOSITION 1: *If development is optimally undertaken, the optimal R&D expenditure plan $y^*(t)$ satisfies*

$$(17) \qquad r/(1 - a) \leq y'(t)/y(t) \leq (r + h)/(1 - a), \qquad 0 < t < \mathrm{T}$$

PROOF:

Logarithmic differentiation of (13) yields

$$(18) \qquad (1 - a)y'/y = r + h - khe^{ht}/(1 + ke^{ht}) = (r + h + rke^{ht})/(1 + ke^{ht})$$

from which (17) easily follows.

Proposition 1 has important conse-

[3]We have assumed that rival preemption renders both the firm's current product and its R&D project worthless. If this assumption were to be relaxed, then the firm must be permitted to reevaluate its options upon rival preemption, considering whether to abandon its R&D project or to continue it, and if to continue, then at what rate. Such possibilities have been considered by the authors (1976b) without a cash constraint.

[4]LEMMA: *If $y(t)$ is a continuous function with $y'(t) > 0$ for $0 \leq t \leq \mathrm{T}$, then* (4) *and* (7) *imply*

(a) $R(t) > 0, \qquad 0 < t < \mathrm{T}$

PROOF:

Suppose R fell to zero before T. Then since π is constant and y is increasing, the right-hand side of (4) would be decreasing so R would fall further, becoming negative. Then R could not increase thereafter, violating (7). Hence (7) can be satisfied only if (a) holds.

FIGURE 1

FIGURE 2

quences. First, the proportionate spending rate for any project undertaken is positive, exceeding the discount rate. The more sharply diminishing returns in R&D, the smaller is a, and the closer the lower bound on y'/y is to r. For an intermediate case of $a = 1/2$, y'/y is no less than twice the discount rate. Second, the cash balance will never be zero before project completion.[5] In particular, a solution in which R&D spending just equals profits cannot be optimal for any interval of time.

Figure 1 illustrates a case in which initial cash is zero and the cash constraint is tight. If the cash constraint is binding, the cash balance will be single peaked,[6] either decreasing throughout or else building up early in the development period while expenditures are low, peaking a bit after spending matches current profits (since $R' = rR > 0$ when $y = \pi$), and then decreasing (concave) to zero.

Third, it is of some interest to note when y'/y achieves its lower limit. It is clear from (18) that if there is no innovational rivalry, ($h = 0$), then $y'/y = r/(1 - a)$. This is true regardless of whether the cash constraint is active or inactive. Thus if the cash constraint is binding, then its effect must be to reduce the absolute spending rate to maintain feasibility, thereby extending the development period. However, the general shape is one of spending rising exponentially at rate $r/(1 - a)$ whether or not the cash constraint is active when $h = 0$.

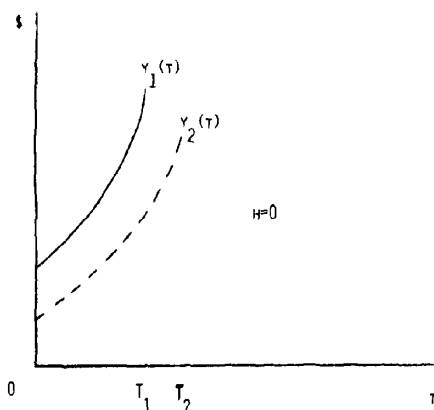



Figure 2 illustrates the effect of the cash constraint; y_1 is the optimal spending program if financing is not binding while y_2 suggests the optimal spending program for the same project if financing must be taken into account. Both paths rise at the same exponential rate, but the constrained path is lower (for comparable times) and of greater duration.

Fourth, note from (18) that if innovational rivals cannot be ignored ($h > 0$), then the proportionate growth rate of spending (as well as its absolute level) is affected by the financing requirement. If the constraint is inactive (so $k = 0$), then, from (18), the spending growth rate achieves its upper limit: $y'/y = (r + h)/(1 - a)$. When the constraint is binding, the proportionate spending growth rate decreases with increasing severity of the constraint (measured by k). An active cash constraint reduces the level and (if $h > 0$) the proportionate growth rate of spending at comparable times and must lengthen the development period, relative to the optimal plan without liquidity problems. Fifth, in case $k = 0$ the proportionate spending growth rate depends only

[5]Since $y' > 0$ throughout the development period, from Proposition 1, it follows from the Lemma of fn 4 that $R(t) > 0$ for $0 < t < T$.

[6]Since $R'' = rR' - y'$, we have $R'' = -y' < 0$ whenever $R' = 0$. Note also that $R'' < 0$ whenever $R' \leq 0$, so R is concave when decreasing.

on r, h, and a and is independent of the merit P of the new project or the difficulty A of achieving it.

III. Cash Constraint Inactive

We consider in detail the case that there is no binding liquidity constraint. Then $k = 0$, and (13)–(16) may be solved[7] to give the optimal expenditure plan

$$(19) \quad y^*(t) = e^{(r+h)t/(1-a)} \cdot [n(r+h)Ab/(1-b)]^{1/a} \qquad 0 \le t \le T^*$$

where

$$(20) \qquad n \equiv a/(1-a)$$

$$(21) \qquad b \equiv 1 - n^{1-a}(r+h)A/[(r+h)P - \pi]^a$$

and development period

$$(22) \qquad T^* = -(\ln b)/n(r+h)$$

provided that

$$(23) \qquad b < 1$$

If (23) does not hold, the project should not be undertaken.

It is evident that the development period (22) is prolonged as either required effort A or current profits π increase and is shortened as the innovational reward P rises. So long as the cash constraint is inactive, the sole impact of current profits is on the attractiveness of the innovation. The larger current profits, the smaller the net gain from innovation (for fixed gross innovational reward P).

The impact of rivalry h and the discount rate r on the development period is ambiguous. One may compute from (22)

$$(24) \quad \partial T^*/\partial h = \partial T^*/\partial r = \frac{1}{n(r+h)^2} \cdot \left[\ln b + \frac{(1-b)}{b}\,\frac{(1-a)P - \pi/(r+h)}{P - \pi/(r+h)}\right]$$

From (23), it is clear that (24) will be negative for relatively modest projects, certainly for $P \le \pi/(1-a)(r+h)$. Thus if, roughly, the new project will not more than double profits ($a = \frac{1}{2}$), then the firm will hasten its development in response to an increase in perceived rivalry or in the discount rate. On the other hand, the authors (1976a) have shown that in the special case of $\pi = 0$, the development period will either be lengthened with increased rivalry (moderately attractive project) or else U-shaped, decreasing with rivalry up to a point and then increasing with further increment in rivalry (better project).

In finding (19)–(23), we assumed the cash constraint would not be tight. This assumption is valid provided that

$$(25) \quad R_0 \ge [(r+h)P - \pi][b^{r/n(r+h)} - b^{1/a}]a/(h+ar) - [1 - b^{r/n(r+h)}]\pi/r$$

as may be verified by substituting from (19) into (4) and integrating, recalling that $R(T^*) \ge 0$ assures $R(t) \ge 0$, $0 \le t \le T^*$. A more intuitive, albeit less comprehensive, condition for the cash constraint to be inactive will be given shortly.

We now show when undertaking new product development is optimal (without a liquidity problem), and calculate its net expected profit improvement in that case.

PROPOSITION 2: *Suppose* (25) *holds; then the R&D project is optimally undertaken if and only if* (23) *holds. Further, if* (23) *and* (25) *hold, the net expected gain from the innovation is*

$$(26) \qquad b^{1/a}[P - \pi/(r+h)] > 0$$

PROOF:

Condition (23) is clearly necessary for T^* in (22) to be positive and finite. To show that it is sufficient, we show that given (23), the maximized value in (5) exceeds (1). Thus it must be shown that

$$(27) \quad \int_0^{T^*} e^{-(r+h)t}(\pi - y^*(t))dt + e^{-(r+h)T^*}P - \pi/(r+h) > 0$$

[7] For instance, put $k = 0$ in (13), substitute the result into (15), and integrate to find expression for λ. One then can eliminate λ from $y(T)$, put the result into (14), and solve for T, (22). With T known, one can find λ and thence y.

where $y^*(t)$, T are given by (19) and (22). But evaluating the left-hand side of (27) reveals it equal to the left-hand side of (26), establishing the sufficiency of (23) (24) for undertaking the *R&D*. Further, since the left-hand side of (27) is the net expected profit improvement from the innovation, the remainder of the proposition follows.

From Proposition 2, when R_0 is large enough, an *R&D* project is both more apt to be undertaken (equation (23)) and will yield a higher net expected gain (equation (26)) when the reward P is high and the current profits π and required effort A are low. Another consequence of Proposition 2 is that product development may be undertaken either in pursuit of improved profits or as a defensive measure against possible losses due to rival entry Without potential rivalry, a new product will be developed only if it is expected to yield greater profits than the current one ($P > \pi/r$). However, a project that would be rejected if there were no fear of rival entry ($p < \pi/r$, $h = 0$) may nevertheless be undertaken if the possibility of such rival preemption were recognized ($h > 0$) and (23) holds. Intuitively, the explanation is that the possibility of rival preemption reduces the expected value of the current product since the expected duration of receipts falls. This enhances the relative value of the new project so it may be undertaken in support of the current line of business as a defensive measure.

The cash constraint will be inactive for a remarkably broad range of projects, as shown in the next result.

PROPOSITION 3: *If*

$$P \leq \pi/a(r + h) \tag{28}$$

then product development can be optimally financed without impedence by the cash constraint.

PROOF:

From Proposition 1, the optimal spending rate rises over the development period. Therefore if $y^*(T^*) \leq \pi$, then surely $y^*(t) \leq \pi$ for all $0 \leq t \leq T^*$, so spending according to (19) is always covered by current receipts. But, from (19) and (22), $y^*(T^*) = n[(r + h) P - \pi] \leq \pi$ provided (28) holds, establishing the proposition.

Proposition 3 indicates, first, that for certain innovations, new product development can proceed without impedence by cash requirements regardless of the effective effort A required or the initial cash balance R_0. If (28) holds and development is worthwhile, then the optimal development schedule is sufficiently leisurely to keep the cash constraint inactive. In other words, because the reward from the new product is modest relative to that of the current product, its development is so prolonged that its difficulty is unimportant. Of course, if the difficulty A is too large, then the project is rejected.

Second, condition (28) indicates that a firm earning high profits from its current product or facing little innovational rivalry (h small) is better able to finance new product development from current profits than one earning low profits or facing intense rivalry. However, a newcomer (not producing the current product and for whom, therefore, $\pi = 0$) nevertheless may develop a superior product more rapidly than the incumbent A newcomer facing the same parameters will choose to develop more rapidly because his potential net innovational reward P exceeds the incumbent's $P - \pi/(r + h)$. However, a newcomer needs a substantial initial cash balance in order for development not to be impeded by a cash constraint. The required initial cash can be determined by setting $\pi = 0$ in (25). An entrant with ample cash will develop faster than an otherwise identical firm already in the market. However, if limited by cash availability, the would-be entrant's speed of development may or may not exceed the incumbent's. The cash required increases with P; a larger reward encourages faster, costlier development. It increases with A for A small, as increased development effort requires more cash. However, for large A, this effect is more than offset by the impact of lessened profitability in reducing the development pace;

cash required then decreases with A.

While the substantial initial cash balance required of the newcomer may pose a barrier to the individual innovator, it need not hamper entry of a firm currently in another line of business. This may help explain why the innovator of a superior product in a particular line of business is often a firm formerly in another line. It also emphasizes the point made earlier that firms currently in a market may not be the only potential innovators of a new product.

Third, since in many instances the expected rewards from the new product will not be more than say twice as high as the expected profits from the current product, the solution with the cash constraint inactive should have considerable applicability. Moreover, the cash constraint can always be rendered inactive by a sufficiently large initial cash balance. Of course (28) is only a sufficient condition that cash not be a constraint; (25) can be satisfied for a far broader range of parameter configurations.

IV. Cash Constraint Active

In case there is no innovational rivalry ($h = 0$) and the cash constraint is binding, the optimal development period T* can be shown[8] to satisfy

$$(29)\quad R_0 + (1 - e^{-rT})\pi/r - A^{1/a}(nr/(e^{nrT} - 1))^{1/n} = 0$$

It follows from (29) that the development period varies directly with A but inversely with R_0 and π. Surprisingly, it is independent of P. The dominant role of π is its effect on cash availability, the same as an increase in R_0. This is the opposite of π's impact when cash was not scarce; then an increase in π served to reduce reward for innovation.

Next we state a counterpart to Proposition 2.

PROPOSITION 4: *If $h = 0$, then a necessary and sufficient condition for the firm to undertake new product development is*

$$(30)\quad P - \pi/r > (nr)^{1/n}A^{1/a}$$

For the proof, see the Appendix.

In the absence of innovational rivalry, a lack of financing (small R_0) may retard development, but, according to Proposition 4, the condition (30) governing whether the project is sufficiently attractive to be undertaken is independent of the financial resources available. In other words, without innovational rivalry, product improvements will never be bypassed solely because of limited cash. The gross profit improvement must be large relative to required effort only.

If the cash constraint is active and there is innovational rivalry, then explicit expression for the three constants T, k, and λ on which y^* depends is not available.[9] However, implicit differentiation of the equations (14)–(16) determining them indicates that, at least for h small, the optimal development period T will be prolonged as the required development effort increases but shortened as either the initial cash R_0 or profits from the current product π increase. In addition, if $h \neq 0$, then a larger reward P hastens development.

Thus it appears that an increase in the expected benefit from innovation P generally hastens development, although there is one interesting exception. If the cash constraint is tight and there is no innovational rivalry, then P does not affect the pace of development; good and excellent projects may proceed equally rapidly just governed by cash availabilities and required effort. Results of our parametric analysis are summarized in Table 1.

V. Summary

We have analyzed the problem of a firm

[8]Substitute (13) into (15) with $h = 0$ to find expression for $a\lambda/(1 + k)$. Use this in (13) and substitute into (16), noting that the coefficient of k in (16) must equal zero since cash is tight.

[9]The difficulty begins with finding neat closed expressions for integrals of the form

$$\int_0^T [e^{(r+h)t}/(1 + ke^{ht})]^n dt$$

TABLE 1

Parameter	Impact on Development Period T* if Cash Constraint is	
	Binding	Not Binding
R_0 = initial cash balance		0
π = current profit rate	[a]	+
P = value of innovation		
A = required innovational effort	+	+
h = intensity of rivalry	+	±

[a]Except 0 if $h = 0$

contemplating new product development. It may anticipate higher profit from an improved product or fear loss of profits were its current product to be displaced by a superior rival product. If, however, a rival product is introduced prior to the firm's own new product, the defense will be unsuccessful and the firm will also lose the resources devoted to development.

We sought an $R\&D$ program to maximize the present expected value of profits, assuming that the firm must finance development entirely from its cash reserves and internally generated profits. A number of interesting findings emerge. First, a cash constraint will not impede development at all for a large class of $R\&D$ projects by established firms, namely those that no more than, say, double the expected gross profits. For such $R\&D$ projects, development can be readily financed from current profits and the cash constraint has no effect. Second, even when the cash constraint is binding, it is never optimal for $R\&D$ spending to just match cash receipts for more than a moment; the cash balance will be single-peaked, becoming zero only upon project completion. Third, in case of no innovational rivalry, the criterion for a project to be worthwhile is independent of the financial requirements of the project and of the financial resources of the firm. No project would be bypassed because of financial limitations. The only possible effect of the cash constraint in this case is to reduce the development pace and prolong the development period. If there may be innovational rivals, an active cash constraint will not only reduce the development pace and prolong the development period of accepted projects, but could also reduce the acceptability of potential $R\&D$ projects. Finally, development was shown to be hastened by a large gross reward and a small required effort, as was expected. In addition, profit from the current product plays two roles; it contributes toward the available cash to finance new product development, but also reduces the attractiveness of introducing a new product that replaces it. The net impact depends on whether the cash constraint is active. So long as financing is not an active constraint, larger current profits retard new product development through their effect on reducing the net gain from innovation. In contrast, if the cash constraint is active, then the role of current profits in providing cash dominates and incremental current profits hasten new product development.

APPENDIX

PROOF of Proposition 4:

If (25) holds with $h = 0$, the conclusion follows immediately from Proposition 2.

Now suppose $h = 0$ and (25) fails. This means (after some manipulation) that

(A1)
$$(R_0 + \pi/r)/P < b^{1/n}[1 - b(1 - \pi/rP)]$$

where b is given by (21) with $h = 0$ there. Note for future reference that (30) is equivalent to

(A2)
$$b < 1 \quad (h = 0)$$

For the project to be worthwhile, it must yield expected rewards that exceed profits available in its absence, i.e.,

(A3)
$$\int_0^{T^*} e^{-rt}(\pi - y)dt + e^{-rT^*}P > \pi/r$$

where T* satisfies (29). We now want to find a condition equivalent to (A3) that can be shown to hold (given (A1)) if and only if (30) does, thereby establishing the proposi-

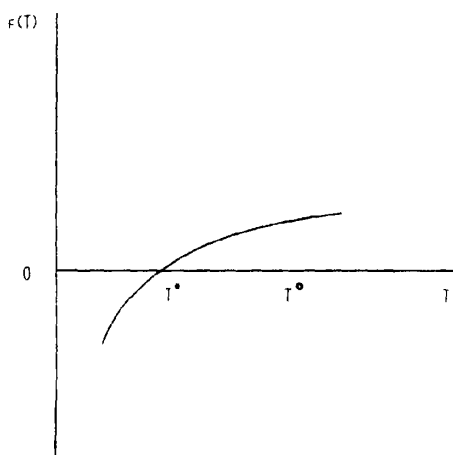



FIGURE 3

tion. Since the cash constraint is tight,

$$\text{(A4)} \qquad \int_0^{T^*} e^{-rt}(\pi - y)dt = -R_0$$

so (A3) can be written as $-R_0 + e^{-rT^*}P > \pi/r$, or equivalently,

$$\text{(A5)} \qquad P/(\pi/r + R_0) > e^{rT^*}$$

Denote the left-hand side of (29) as $f(T)$. Then $f'(T) > 0$, $\lim_{T\to 0} f(T) < 0$, and $\lim_{T\to\infty} f(T) > 0$.

Define T* by

$$\text{(A6)} \qquad e^{rT^*} = P/(\pi/r + R_0)$$

Now $T^0 > T^*$ if and only if $P/(\pi/r + R_0) = e^{rT^0} > e^{rT^*}$; it is clear from this and Figure 3 that (A5) holds if and only if $f(T^0) > 0$. Recalling definition (A6), it can be shown after some manipulation that $f(T^0) > 0$ if and only if

$$\text{(A7)} \qquad 1 - (1 - b)^{1/(1-a)} > ((R_0 + \pi/r)/P)^a$$

We have now shown that (A3) and (A7) are equivalent (given (A1)). It must be shown that (given (A1)), (A7) implies (30) and (30) implies (A7) to complete the proof. But if (A7) holds, then (A2) does and hence so does (30). We show that (30) implies (A7) by the following:

$$\text{(A8)} \quad ((R_0 + \pi/r)/P)^a < b[1 - b(1 - \pi/rP)]^a < b < 1 - (1 - b)^{1/(1-a)}$$

The first inequality follows from (A1); the second follows since the square-bracketed term is less than one, by (30); and the third is true since

$$[1 - (1 - b)^{1/(1-a)}] - b = (1 - b)[1 - (1 - b)^a] > 0$$

Hence (A7) follows, completing the proof.

REFERENCES

Kenneth J. Arrow and Mordecai Kurz, *Public Investment, The Rate of Return, and Optimal Fiscal Policy*, Baltimore 1970.

A. E. Brown, "New Definitions for Industrial R&D," *Res. Manage.*, Sept. 1972, *15*, 55–57.

M. I. Kamien and N. L. Schwartz, "Timing of Innovations under Rivalry," *Econometrica*, Jan. 1972, *40*, 43–60.

——— and ———, "Risky R&D with Rivalry," *Annals Econ. Soc. Measure.*, Jan. 1974, *3*, 267–77.

——— and ———, "Market Structure and Innovation: A Survey," *J. Econ. Lit.*, Mar. 1975, *13*, 1–37.

——— and ———, (1976a) "On the Degree of Rivalry for Maximum Innovative Activity," *Quart. J. Econ.*, May 1976, *40*, 245–60.

——— and ———, (1976b) "Potential Rivalry, Monopoly Profits and the Pace of Inventive Activity," *Rev. Econ. Stud.*, forthcoming.

Edwin Mansfield et al., *Research and Innovation in the Modern Corporation*, New York 1971.

Frederic M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

Bargaining Theory, Wage Outcomes, and the Occurrence of Strikes: An Econometric Analysis

By HENRY S. FARBER*

The purpose of this study is to explore and test with micro-economic data a simple model of trade union wage determination that is explicitly derived from a naive, but instructive, model of collective bargaining. Although the importance of trade union wage behavior is often recognized in studies of aggregate wage inflation, it is well known that not much progress has been made in assimilating micro-economic analyses of bargaining into such models.[1] The incorporation of explicit bargaining structures should lead to a better understanding of how wage changes and strike frequencies in the union sector are influenced by changes in the economic environment, as well as a better understanding of the behavioral processes of collective negotiations.

Section I of this paper contains the development of the model to be tested, while Section II provides an interpretation of the model and spells out the testing framework. Section III contains a description of the sample, while Section IV contains the empirical results and an example of their use in analyzing a particular bargaining situation. Section V provides a summary of the results and some conclusions drawn from the analysis.

I. The Model

The model of wage determination used here follows the model developed by Orley Ashenfelter and George Johnson to analyze strike activity. The basic assumption of this model is that the firm attempts to maximize its present value while faced with a tradeoff between the size of the wage increase demanded and the length of the strike incurred. The shape of this tradeoff is determined by a "concession schedule" which denotes the minimum wage increase acceptable to the union rank and file after a strike of a given length. This schedule has a negative slope, reflecting the rate at which the rank and file reduce their expectations of a wage increase in response to hardships imposed on them by a strike and to "new" information learned from a strike about the degree of employer resistance to union wage demands.[2] In simple terms, they become willing to settle for less as the strike progresses. The role of the union leadership is to convey to the management the shape of the concession schedule as well as to provide information to the rank and file regarding feasible wage demands.[3] The concession schedule can be written as

$$Y_A = g(s) \tag{1}$$

$$dY_A/ds = g'(s) < 0 \tag{2}$$

where Y_A is the minimum acceptable proportionate wage increase after a strike of length s. This function is shown as the convex curve in Figure 1.

Although not necessarily convex, it is convenient to assume that the concession

*Assistant professor of economics, Massachusetts Institute of Technology. I wish to thank Orley Ashenfelter for helpful discussions throughout the preparation of this study, and the referee for useful comments. Financial support for this research was provided from a grant to the Princeton University department of economics from the Sloan Foundation.

[1] See George de Menil for an example of an attempt at such a synthesis.

[2] It is important to mention that this concession schedule is not based on any hypothesis of maximizing behavior on the part of the rank and file. Therefore, the concession schedule may not reflect economically rational choices of wage increase and strikes by the workers. While any discussion of gains or losses to the workers from strikes must necessarily be imprecise, it is probably true that strikers rarely are net monetary gainers from work stoppages and, hence, they are not acting rationally when they strike.

[3] See Arthur Ross, pp. 1-74, and Ashenfelter and Johnson, pp. 36-37, for more detailed discussions of this concept of internal union behavior.

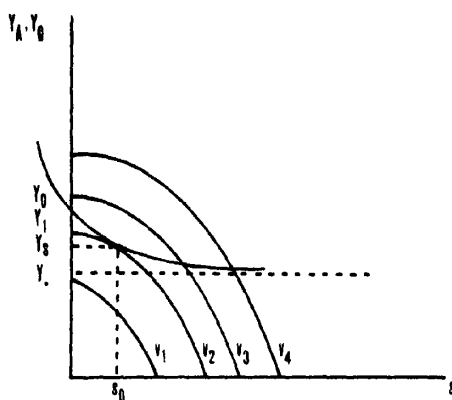



FIGURE 1

schedule has this property. This assumption seems reasonable despite the well-known diagram of John Hicks, p. 143, of a similar relationship which has a concave section. The notion of convexity is based on the idea that there is some irreducible minimum wage increase, perhaps negative, that the union will be willing to strike for indefinitely and that reflects the best alternative wage available with certainty to union members.

Since the present value of the firm is simply the discounted value of its profit stream, maximizing present value implies that the firm chooses the optimal tradeoff between foregone profits during a strike and increased labor costs after a strike. The present value of the firm (v) can be written, *ceteris paribus*, as

$$v = f(Y_g, s) \tag{3}$$

where Y_g is the wage increase granted and both $\partial v/\partial Y_g$ and $\partial v/\partial s$ are negative. The tradeoff can be illustrated by totally differentiating (3) to get

$$dv = \frac{\partial v}{\partial Y_g}\, dY_g + \frac{\partial v}{\partial s}\, ds \tag{4}$$

and then setting $dv = 0$ to find

$$\frac{dY_g}{ds} = -\frac{\partial v}{\partial s}\bigg/\frac{\partial v}{\partial Y_g} < 0 \tag{5}$$

which is an expression for the slope of the firm's iso-value curves. A family of these curves are drawn in Figure 1. Clearly, $v_1 > v_2 > v_3 > v_4$. The firm's goal is to reach the iso-value curve corresponding to the highest discounted present value or, in other words, to get as close to the origin in Figure 1 as possible. However, the firm is constrained by the union's concession schedule (1).

The optimum decision for the firm is to settle at the point of tangency between the union concession schedule and an iso-value curve. The solution results in a proportionate wage increase of Y_s and a strike of length S_0. It is possible that the iso-value curves are everywhere steeper than the concession schedule. In this case the firm's optimal decision is to settle without a strike and grant a proportionate wage increase of Y_0.

The final settlement in this model is not Pareto optimal, as is clearly illustrated in Figure 1. The final settlement occurs at Y_s, but a strike of length s_0 takes place resulting in both lost income to the members of the union and a reduction in the present value of the firm. If the firm could have settled for any wage increase up to Y_1 without a strike, it would have been on a higher iso-value curve. The union members would also have won at least the same wage increase with no lost income. The intransigence of the union membership, as opposed to the leadership, is the source of this result.[4]

Parameterizations of the union concession schedule and the firm's present value function that have been suggested are

$$Y_A = Y_* + (Y_0 - Y_*)e^{-as} \tag{1'}$$

[4]The implication of a good deal of the bargaining literature is that the occurrence of a strike generally indicates either that one (or both) of the parties was "irrational" or that a failure of communication occurred. No attempt is made in this literature to analyze either the causes or consequences of disputes, and the question of wage determination is considered separately. Hicks, pp. 144–47, Frederik Zeuthen, pp. 104–21, and Robert Bishop take this position. The unique feature of the model developed here is that wage outcomes and the occurrence of strikes are codetermined through a mechanism combining optimizing behavior on the part of the firm with political behavior by the members of the union.

and[5]

$$(3')\quad v = \int_{s}^{\infty} [PQ - LW(1 + Y_g)]e^{-rt}dt - \int_{0}^{\infty} He^{-rt}dt$$

Integration of (3′) yields

$$(6)\quad v = [PQ - LW(1 + Y_g)]\frac{e^{-rs}}{r} - \frac{H}{r}$$

In (1′) $Y^* = g(\infty)$, $Y_0 = g(0)$, and a represents the rate of decay of the union's wage demands. In (6) PQ is total revenue, L is labor input in man-hours, W is the previous wage rate, H is fixed costs, and r is the rate at which the firm discounts future earnings.

Solving (6) for Y_g yields

$$(7)\quad Y_g = \frac{PQ}{LW} - \frac{(vr + H)}{LW}e^{rs} - 1$$

The optimal solution for the firm is where

$$(8)\quad \frac{dY_A}{ds} = \frac{dY_g}{ds}$$

Equating the derivatives with respect to s of (1′) and (7) and solving for $Y_s = Y_g = Y_A$ yields

$$(9)\quad Y_s = \left(\frac{a}{a + r}\right) Y_* + \frac{r}{a + r}\left(\frac{1}{S_L} - 1\right)$$

where $S_L = WL/PQ$ or labor's share of total sales prior to the latest negotiations.

Equation (9) is valid only for those negotiations where a strike occurs. Where there is no strike, a corner solution occurs and $Y_s = Y_0$, rather than the relationship defined in (9). The fact that the concession schedule is negatively sloped and everywhere flatter than the iso-value curves implies that

$$(10)\quad Y_s = Y_0 < \left(\frac{a}{a + r}\right) Y_* + \frac{r}{a + r}\left(\frac{1}{S_l} - 1\right)$$

where no strike occurs. This inequality will prove useful in formulating an estimation framework.

[5]These are suggested by Ashenfelter and Johnson, pp 37 38. Clearly, there are other possibilities for the functional forms, and it would be interesting to explore them.

II. The Estimation Framework

The equality in (9) and the inequality in (10) express the wage settlement as a function of parameters (a, Y_*, r) which, although they have straightforward economic interpretations, are not directly observable. It seems natural to assume for the present that these are to be estimated from data on Y_s, S_L, and the occurrence or nonoccurrence of a strike in a particular bargaining situation.

To be explicit, assume that (9) and (10) have an additive error distributed as $N(0, \sigma^2)$ where σ^2 is an unknown variance. This yields

$$(11a)\quad Y_{si} = \frac{a}{a + r} Y_* + \frac{r}{a + r}\left(\frac{1}{S_{Li}} - 1\right) + \epsilon_i$$

If a strike occurred in observation i and

$$(11b)\quad Y_{si} < \frac{a}{a + r} Y_* + \frac{r}{a + r}\left(\frac{1}{S_{Li}} - 1\right) + \epsilon_i$$

if there was no strike in observation i. Let

$$(12)\quad Z_i = Y_{si} - \left[\left(\frac{a}{a + r}\right)Y_* + \frac{r}{a + r}\left(\frac{1}{S_{Li}} - 1\right)\right]$$

Rearranging terms and substituting equation (12) into relations (11a) and (11b) yields

$$(13)\quad \epsilon_i = Z_i$$

$$(14)\quad \epsilon_i > Z_i$$

for the strike and no-strike cases, respectively.

In this situation maximum likelihood estimation implies that we maximize the probability of occurrence or nonoccurrence of strikes with given wage outcomes in the sample. Relations (12), (13), and (14) along with assumption of independence of the ob-

TABLE 1 - FIRMS AND UNIONS IN SAMPLE

Number	Firm	Union
1	American Cyanamid	International Chemical Workers Union
2	Firestone Tire and Rubber Company	United Rubber, Cork, Linoleum, and Plastic Workers of America
3	General Electric	International Union of Electrical, Radio, and Machine Workers
4	General Motors	United Automobile, Aircraft, and Agricultural Implement Workers of America
5	International Paper	International Brotherhood of Pulp, Sulphite, and Paper Mill Workers, United Papermakers and Paper Workers[a]
6	PPG Industries[b]	United Glass and Ceramic Workers of North America
7	Simmons	United Furniture Workers of America
8	Sinclair Oil Corporation	Oil, Chemical and Atomic Workers International Union[c]
9	United States Steel	United Steelworkers of America
10	Weyerhaeuser Timber Company	International Woodworkers of America

[a] Prior to 1956 *UPP* was International Brotherhood of Papermakers.
[b] Previously Pittsburgh Plate Glass.
[c] Prior to 1955 *OCAWIU* was Oil Workers International Union.

servations imply a likelihood function of the form

$$(15) \qquad L = \prod_{i=1}^{m} f(Z_i) \prod_{i=m+1}^{n} (1 - F(Z_i))$$

where $f(Z_i)$ is a normal probability density function and $F(Z_i)$ is a cumulative normal distribution function.[6] It is clear that the maximum likelihood estimates of a, Y_*, and r from (15) cannot be computed analytically, and for this reason numerical methods will be used.[7]

III. The Sample

The sample is composed of negotiated wage settlements for ten large manufacturing firms in ten different 2-digit *SIC* industries.[8] They are listed in Table 1 along with the international union with which they negotiate. In the two instances where a firm negotiated with more than one union the negotiations were held jointly. There are a total of eighty contracts in the sample for the ten firms covering the period from 1954 to 1970. Strikes occurred in twenty-one of the cases.

The dependent variable Y_s is the average annual percentage change in the negotiated base wage rate for janitors computed between contracts.[9] This is the conventional base wage over which most negotiations take place. The only other observable variable in (9) is labor's share of total sales, S_L. Data were collected on total labor costs for each firm and on total net sales for each firm for the fiscal year preceding each settlement. Total labor costs include fringe benefits such as payments into insurance and pension funds as well as wages and salaries of both union and nonunion workers.[10]

[6] The term $f(\epsilon_i)$ represents the probability density of the occurrence of a strike with given wage outcome characteristics. The probability of nonoccurrence of a strike is equal to the probability that inequality (14) holds. This can be represented by $1 - Pr(\epsilon_i < Z_i)$ or $1 - F(Z_i)$. Strikes occurred in the first m observations.

[7] See Takeshi Amemiya for a discussion of estimation where the distribution of the dependent variable is truncated normal. In the context of a linear model he proves the consistency of maximum likelihood estimators.

[8] By choosing firms in different industries we are assured that there is a degree of independence among the observations. The existence of wage patterns within industries in manufacturing implies that wage settlements reached by firms in the same industry are not independent. Therefore, it would be incorrect to increase the number of observations by including other firms from the same industries.

[9] I am grateful to Daniel Hamermesh for making these data available to me. For further information concerning the construction of the wage series and the incorporation of automatic cost-of-living increases and deferred increases, see Hamermesh, pp. 505-06.

[10] Because these data are all that are available it must be assumed that fringe benefits increase at the same rate as wages. The inclusion of nonunion labor costs causes little distortion in the firm's estimates of the costs of any wage increase because, in general, any wage increase granted to production workers is even-

IV. Specification of the Testing Model

The assumption that a and Y_* are constant for every firm in every year is clearly not tenable, and it will not be maintained in the empirical work that follows. In its place hypotheses will be formed concerning the exogenous determinants of a and Y_*, and these will be tested using the formulation described above. It is not possible to estimate both a and r independently because equation (9) is homogeneous of degree zero in those parameters, and, for this reason, it will be assumed that the rate of discount (r) is .1 in order to identify a.

The rate of concession of the union (a) is a function of two factors. The first is the ability of the union members to replace income lost during a strike through such things as strike benefits or alternative employment opportunities The second is the militancy or mood of the rank and file as determined by such things as previous changes in real wages and how effective they can expect a strike to be in terms of imposing costs on their employer. Anything which increases the militancy of the rank and file or reduces the impact of income lost during a strike will reduce the union's rate of concession. This hypothesis is very general and empirical exploration is necessary in order to more precisely ascertain the determinants of a.

A preliminary specification for a is

$$(16) \quad a_{it} = \alpha_0 + \alpha_1 \cdot FB_{it} + \alpha_2 \cdot U_t + \alpha_3 \cdot \pi_{it-1} + \alpha_4 \cdot PC_t + \alpha_5 \cdot S_{Lit} + \alpha_6 \cdot DRW_{it}$$

where

a_{it} = rate of concession of union i in year t

FB_{it} = union fund balances per member of union i in year t

U_t = unemployment rate in year t

π_{it-1} = rate of return on assets for firm i in year $t - 1$

PC_t = dummy variable having a value of one if the observation is in the period from 1962–66, and having a value of zero otherwise

S_{Lit} = labor's share of total sales in firm i in year t

DRW_{it} = average annual rate of change of real wages over the life of the previous contract in firm i. This is computed as the difference between the previous average annual rate of change in wages and the average annual rate of change in the consumer price index over the life of the previous agreement.

Union fund balances per member (FB) are a proxy for the ability of the union to pay strike benefits to its members. However, the measure is faulty in that it does not reflect actual benefits paid, and it reflects only balances in the funds of the international unions while neglecting assets of the local unions.[11] Nevertheless, it is the best measure available. Higher strike benefits should reduce the impact of lost income during a strike and lower the rate of concession. Thus, it is expected that $\alpha_1 < 0$.

The national unemployment rate for the month of negotiation (U) is hypothesized to vary inversely with the availability of temporary employment for striking workers. Temporary employment is another way of reducing the impact of income lost during a strike. Thus, the unemployment rate is expected to vary directly with the rate of concession, and it is expected that $\alpha_2 > 0$.

Past profits of the employer (π_{t-1}) are expected to affect worker militancy directly, and lower the rate of concession. It is hypothesized that $\alpha_3 < 0$.

When wage guidelines (PC) are in effect,

tually extended to all employees. See Albert Rees, p. 73. In the simple model used here it is assumed that union labor costs are the only costs that are avoided during a strike. Thus the use of total labor costs leads to an overestimate of the costs avoided during a strike, but constraining all other costs, such as expenditures on raw materials, to be fixed results in an underestimate of these avoided costs. These distortions are at least partially offsetting, and they are ignored here

[11]These data were collected from the convention proceedings and newspapers of the various unions in the sample

even if they are of the voluntary variety of the 1962–66 period, outside pressure is brought to bear on the union to concede. Thus, it is expected that the presence of guidelines will increase the rate of concession and that $\alpha_4 > 0$.

Labor's share of total sales (S_L) is hypothesized to be directly related to the ability of the union to cause a cessation of production in the short run. A union which could not shut down the operations of its employer would find its members becoming discouraged very quickly, and it would concede rather rapidly. Thus, it is expected that $\alpha_5 < 0$. It is interesting to contrast this to the role of labor's share in the direct formulation of (9). There S_L has a negative effect on the size of the wage settlement because if labor's share is *large* then the firm's optimal tradeoff between foregone profits during a strike and decreased future labor costs will be skewed toward the latter. The result will be a higher propensity to strike and a lower wage settlement. On the other hand, the effect of a large share for labor through the rate of concession is to lower the rate of concession making a strike less attractive to the employer and, hence, yielding a higher wage settlement.

The lagged rate of change of real wages (DRW) is expected to have an inverse relationship with worker militancy. It is hypothesized that when real wages are rising slowly or falling, workers will be more militant and concede more slowly. Thus, it is expected that $\alpha_6 > 0$.

There is one other very clear prediction about the rate of concession that can be inferred from the model: the rate of concession must be positive for the model to make any sense. After estimates of the α's are obtained in the next section, they will be used to obtain predicted values for the rate of concession for each observation in order to see if this condition is met.

The minimum acceptable wage increase after an "infinitely" long strike (Y_*) can be interpreted as that proportionate change in wages which would make the union wage equal to the best alternative wage available to the union members;

$$(17) \qquad Y_* = \frac{W_A - W_U}{W_U}$$

where W_A equals the best alternative wage and W_U equals the union wage.[12] If the alternative wage is interpreted as the average nonunion wage in manufacturing, then $-Y_*$ is a similar concept to the union-nonunion wage differential measured by H. Gregg Lewis, pp. 219–21, who argued that union wages were rigid in that they did not respond to cyclical changes in the economy. Thus, if the alternative wage is not rigid, Y_* will be directly related to cyclical changes. In other words, if the economy is strong and if there is a high demand for labor, the alternative wage will rise relative to the union wage and Y_* will increase. The unemployment rate (U) is used here as a measure of the state of the labor market, and it is expected that Y_* will be inversely related to the unemployment rate.

There is obviously some interindustry variation in wage differentials, and from (17) it is easily seen that Y_* will be inversely related to such variations. However, an investigation of the causes of these variations is beyond the scope of this study. In order to account for these differences, industry dummy variables will be used in the estimation of Y_*.

The following specification is assumed:

$$(18) \qquad Y_{*it} = B_0 U_t + \sum_{i=2}^{10} B_i D_i$$

where Y_{*it} is the minimum acceptable wage increase in year t for union i and D_i is a dummy variable having a value of one in industry i and zero elsewhere.[13] The B_i can be interpreted as the difference between the minimum acceptable wage increase in firm i

[12]Equation (17) is obtained from the assumption made that $(1 + Y_*)W_U = W_A$

[13]It is not possible to estimate coefficients for dummy variables in all ten firms because U_t shows no cross-sectional variation, and, hence, acts as a constant term. The result is that all ten dummy coefficients are not identified and one falls into a rather unique sort of "dummy variable trap." A solution is to set the dummy variable for industry 1 equal to zero and to measure all other Y_{*i} relative to firm 1

and that in firm 1 at any point in time. Firms in industries where the union-non-union wage differential is relatively large are expected to have relatively low values for their B_i. It is expected that B_0, the coefficient of U_t, will be negative.

The increase Y_* must satisfy three constraints if the interpretation of the model presented here is to make sense. First, $Y_* > -1$. In other words, the minimum *acceptable* wage *rate* must be positive Second, $Y_* \leq Y_s$. The minimum acceptable wage change must be less than or equal to the settlement finally agreed upon. Third, given our interpretation of Y_* and the fact that the union wage differentials are generally positive, it is expected that the predicted values of $Y_* \leq 0$ This is a stronger constraint than the second listed above because $Y_s \geq 0$ for all observations in the sample.

The qualitative expectations concerning the parameters a, and Y_* are summarized below.

$$\alpha_1 < 0 \qquad \alpha_4 > 0 \qquad B_0 < 0 \qquad Y_* \leq Y_s$$
$$\alpha_2 > 0 \qquad \alpha_5 < 0 \qquad a > 0 \qquad Y_* \leq 0$$
$$\alpha_3 < 0 \qquad \alpha_6 > 0 \qquad Y_* > -1$$

V. Empirical Results

Estimates of the parameters of equations (16) and (18) are obtained by substituting (16) and (18) into (12) This is in turn substituted into (15) which is then maximized numerically yielding maximum likelihood estimates of the α_i and B_i These estimates are distributed asymptotically normally Estimates of their asymptotic standard errors are derived using elements of the inverse second derivative matrix of the *log* likelihood function.[14] The parameter estimates along with their standard errors are presented in Table 2.

To facilitate the discussion of the results we define a group of functions of a and Y_* The first is R_T or the concession range,

$$R_T = Y_T - Y_* \tag{19}$$

[14]These derivatives are symmetric numeric derivatives For further information concerning the approximation used see Stephen Goldfeld and Richard Quandt, pp 18 20.

TABLE 2—ESTIMATES

Equation (16) – a	$\hat{\alpha}$	Asymptotic Standard Error
Constant	4.141[a]	.8666
FB	0075	.0066
U	.0262	.0729
π	- 1582	1.470
PC	.3137[a]	.1826
S_L	–8 579[a]	2 469
DRW	.8129	2.314
Equation (18) – Y_*	$\hat{B}$	
U	–.0095[a]	0056
Firm 2	0018	0116
Firm 3	- .0428	0370
Firm 4	0093	0186
Firm 5	0093	0141
Firm 6	- 0072	.0147
Firm 7	– 0120	0143
Firm 8	– 0924[a]	0276
Firm 9	– 0607[a]	0197
Firm 10	0094	0144

[a]Signifies that the coefficient is asymptotically significantly different from zero at the 5 percent level on the basis of a one-tailed test where appropriate, or a two-tailed test where the former is not appropriate

This is the maximum amount by which the firm may reduce through a strike the proportionate wage increase it grants to its workers. The second is R_F or the range of future concessions which is defined as

$$R_F(s) = Y_A(s) - Y_* \tag{20}$$

Since Y_A is a function of the length of strike incurred, R_F represents the amount of the concession range (R_T) still to be conceded after a strike of length s. Clearly, $R_F(0) = R_T$ Next, we define T as the half-life of $R_F(s)$ or, in other words, as that length of strike necessary to reduce $R_F(s)$ to one-half of its original value (R_T). Solving

$$\frac{R_F(T)}{R_T} = \frac{R_F(s+T)}{R_F(s)} = e^{-at} = .5 \tag{21}$$

for T yields

$$T = ln\,(2)/a \tag{22}$$

T is independent of s, and it represents the further length of strike required to reduce $R_F(s)$ to $.5R_F(s)$ for any s. The R_T and T are interesting concepts because they represent,

TABLE 3

Firm	$\bar{Y}_s$	$\bar{Y}_0$	$\bar{Y}_*$	$\bar{R}_T$	$\bar{a}$	$\bar{T}$	$\bar{\Delta}$
1. American Cyanamid	.0516	.0581	−.0472	.1054	1.99	.351	0495
2. Firestone	.0504	.0619	−.0480	.1099	2.16	.325	.0504
3. General Electric	.0506	.0679	−.0872	.1552	.969	.726	.0955
4. General Motors	.0460	.0514	−.0408	.0933	2.21	.315	.0425
5. International Paper	.0459	.0470	−.0405	.0876	2.32	.299	.0422
6. PPG	.0360	.0466	− 0530	.0991	1.54	.462	.0559
7. Simmons	.0520	0588	− 0606	1194	1.59	.444	.0645
8. Sinclair Oil	0391	0495	−.1368	.1859	3.31	.210	.1584
9. U.S. Steel	.0391	0473	−.1048	1522	772	.976	.1170
10. Weyerhaeuser	0391	.0443	−.0581	.1024	1.92	.364	.0616
AVERAGE	0451	.0531	−.0670	.1204	2.00	.415	.0731

$\bar{Y}_s$ = proportionate wage increase granted
$\bar{Y}_0$ = minimum acceptable wage increase with no strike
$\bar{Y}_*$ = minimum acceptable wage increase after infinite strike
$\bar{R}_T = \bar{Y}_0 - \bar{Y}_*$ = concession range
$\bar{a}$ = rate of concession
$\bar{T}$ = half-life of concession range in years
$\bar{\Delta} = -\bar{Y}_*/(1 + \bar{Y}_*)$
= union-nonunion wage differential

respectively, to the firm the total possible gains from a strike and how long it will take to achieve a fixed proportion of those gains.

Values of R_T for each observation in the sample were computed by obtaining estimates for Y_0 and subtracting the predicted value of Y_*. Estimates of Y_0 were obtained by solving equation (1′) for Y_0 using the predicted values for Y_* and a as well as data on Y_s and s. The T were computed from equation (22) using the estimates obtained for a. Presented in Table 3 are average values of Y_0, R_T, and T for each firm as well as their overall averages.

By accepting the interpretation of Y_*, given in equation (17), as a function of the union-nonunion wage differential, we computed the standard union-nonunion wage differential (Δ) as a function of Y_*.

$$\Delta = \frac{W_U - W_A}{W_A} = \frac{-Y_*}{1 + Y_*} \qquad (23)$$

The differential computed from the average predicted Y_* for each firm is presented in Table 3. The overall average differential estimated in this study of .0731 is somewhat smaller than the differentials estimated by Lewis using aggregate earnings data. The fact that the concept of the best alternative wage used in this study is not strictly the nonunion wage but is some weighted average of union and nonunion wages may account for the difference in estimates.

All of the estimated coefficients in the equation determining a except that for union fund balances (FB) have the expected sign, but only those for the wage guidelines dummy (PC) and labor's share (S_L) are significantly different from zero. The nonsignificance of the variables representing "militancy" (π, DRW) and "income replacement" (FB, U) suggests that these variables have little explanatory power. However, by the same criteria, those reflecting outside pressure (PC) and potential strike effectiveness (S_L) do seem to play a role in the determination of the rate of concession. Further empirical work is necessary in order to sort out more precisely the effects of these variables on the wage settlement and strike length.

Predicted values of a and their asymptotic standard errors were computed for each observation, and, in accordance with our expectations, they are all significantly positive at the 5 percent level. Average values of a for each firm are presented in Table 3.

The average half-life (T) of the concession ranges (R_T) in Table 3 is .415 years or approximately five months. This is longer than any strike in the sample. In fact, in none of the bargaining situations in the sample did the employer stand a strike long enough to gain half of the potential reduction in the wage increase.

The significance of the coefficient of the unemployment rate in the equation deter-

mining Y_* implies that the minimum acceptable wage increase moves cyclically, providing support for the wage rigidity hypothesis. Two of the firm dummies have coefficients significantly different from zero implying that for a given unemployment rate these firms face a significantly different minimum acceptable wage increase (Y_*) than does firm 1.[15] The fact that these coefficients are negative implies that the union-nonunion wage differentials in the industries containing the two firms (petroleum and steel) are larger than that in the industry of firm 1 (chemicals). The average differentials (Δ) in Table 3 are, in fact, much larger for the firms in these industries than for American Cyanamid.[16]

As was done with the rate of concession, predicted values and asymptotic standard errors were computed for Y_* for each observation. Average values of Y_* for each firm are presented in Table 3. Once again the results are in accordance with our expectations. All of the predicted Y_* were significantly greater than -1 at any reasonable level. They were all also negative, although only fifty-two of the eighty were significantly negative at the 5 percent level. However, seventy-nine of the eighty were significantly negative at the 10 percent level.

It is illuminating to examine a particular bargaining situation in the context of the empirical results. The strike against the Firestone Tire and Rubber Company by the United Rubber Workers in 1967 lasted ninety-one days or .25 years. This was the longest strike in the sample relative to the half-life of the concession range although it is not the longest strike in absolute terms. The ratio of the strike length to T was .758. The final settlement was for an average annual proportionate wage increase of 4.61 percent. The wage increase the firm would have had to offer to avert the strike (Y_0) was computed to be 10.14 percent. On the basis of these figures the strike seems to have significantly reduced the size of the wage increase granted.

Using the financial data of the firm it is possible to compute in very rough terms the benefit-cost ratio of the strike to Firestone. The ratio of net income to net sales, and the ratio of wages, salaries, and employee benefits to net sales were fairly stable at .055 and .3, respectively, for the period surrounding 1967.[17] An approximation to the benefits from the strike is

$$(24)\quad BEN = \int_s^{\infty} [(1 + Y_0) - (1 + Y_s)].3PQe^{-rt}\,dt$$

This is the reduction in future labor costs resulting from granting a wage increase of Y_r rather than Y_0. The approximate loss can be expressed as

$$(25)\quad LOS = \int_0^{s} .055PQe^{-rt}\,dt$$

or the profits foregone during the strike. The BEN/LOS ratio is computed using the values $r = .1$, $s = .25$, $Y_r = .0461$, and $Y_0 = .1014$. The resulting ratio is approximately 11.9, and it is so large as to leave little doubt as to Firestone's wisdom in taking the strike.

It is also interesting to compute the wage increase (Y_1) Firestone could have granted without a strike while remaining on the same iso-value curve it settled on after the strike. The Y_1 will be less than Y_0 and greater than Y_s, but its position in this range has implications for union policy.[18] The Y_1 is computed by equating the gain from averting the strike (foregone profits) with the cost of averting the strike (increased future labor costs) and solving for Y_1. Foregone profits are given in equation

[15] A likelihood ratio test of the null hypothesis that all of the firm dummies in the equation determining Y_* are equal versus the alternative that they vary by firm decisively rejects the null hypothesis at any reasonable significance level.

[16] It is interesting to note that, while the implied wage differentials in the petroleum and steel industries are the largest in the sample, the average concession rate of Sinclair Oil is the largest in the sample and the average concession rate of United States Steel is the smallest. This raises some interesting questions concerning the relationship between the size of the long-run union-nonunion wage differential, the ability of the union to win a wage increase at a point in time, and the propensity to strike in an industry

[17] See Firestone Tire and Rubber Company, pp. 8-9. The calculations that follow are not entirely consistent, but they are useful for expository purposes.

[18] See Figure 1.

(25) while increased future labor costs (COS) are

$$(26)\quad COS = \int_s^\infty [(1 + Y_1) - (1 + Y_s)].3PQe^{-rt}dt$$

The solution yields $Y_1 = .0507$. This is much closer to Y_s than to Y_0. This suggests that Firestone's iso-value curves, shown in Figure 1, were quite flat, and even a substantial moderation in Y_0 by the union would not have averted the strike.[19]

VI. Conclusions

In the previous section the bargaining model derived earlier was estimated in order to determine the reasonableness of the formulation. On the basis of these results it can be conditionally concluded that the model is indeed reasonable. The empirical results lend support to virtually all of our a priori qualitative hypotheses concerning the structure of the model.

More specifically, it was found that the rate of concession of the union members is responsive both to the outside pressure of wage guidelines and to the potential effectiveness of a strike. Another finding was that the minimum acceptable wage change after a long strike can tentatively be modeled as that change which makes the union wage equal to some alternative wage. In this context the hypothesis of the rigidity of union wages was tested and could not be rejected.

The results of this study demonstrate the feasibility of analysing the outcome of collective bargaining using a simple model. The ease of interpretation of the parameters add considerable richness to the empirical results, and ultimately, may even be of use in micro-economic decision making.

[19]A rational firm would be willing to settle (unexpectedly) without a strike for anything less than Y_1 In fact, its final offer should approximate Y_1 Firestone's final public offer was .0391 This is less than Y_1, probably because common sense bargaining strategy dictates that when the parties are not close to agreement they do not publicly concede all that they are actually willing to give. See U.S. Bureau of Labor Statistics, pp. 3 4, for a description of the negotiations. The numerical estimates of Y_1 as well as the estimates of the benefit-cost ratio depend crucially upon the assumed value of the rate of discount. However, the qualitative results are not affected. An increase in the assumed discount rate will increase the value of profits foregone relative to future savings in labor costs, and the result will be a decrease in the benefit-cost ratio However, because of the manner in which Y_0 is computed, an increase in *a ex post* will increase our estimate of Y_0. This will increase the benefit-cost ratio and offset the first effect An increase in the rate of discount will increase our estimate of Y_1, but because our estimates of Y_0 is also increased the relative position of Y_1 in the $Y_0 - Y_s$ range will not be substantially changed. These assertions were verified using values of r between 05 and .5.

REFERENCES

T. Amemiya, "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, Nov. 1973, *41*, 997-1016.

O. Ashenfelter and G. E. Johnson, "Bargaining Theory, Trade Unions, and Industrial Strike Activity," *Amer. Econ. Rev.*, Mar. 1969, *59*, 35-49.

R. L. Bishop, "A Zeuthen-Hicks Theory of Bargaining," *Econometrica*, July 1964, *32*, 410-17.

George de Menil, *Bargaining: Monopoly Power Versus Union Power*, Cambridge, Mass. 1971.

Stephen M. Goldfeld and Richard E. Quandt, *Nonlinear Methods in Econometrics*, Amsterdam 1972.

D. Hamermesh, "Wage Bargaining, Threshold Effects, and the Phillips Curve," *Quart. J. Econ.*, Aug. 1970, *84*, 501-17.

John R. Hicks, *The Theory of Wages*, New York 1963.

H. Gregg Lewis, *Unionism and Relative Wages in the United States*, Chicago 1963.

Albert Rees, *The Economics of Trade Unions*, Chicago 1962.

Arthur M. Ross, *Trade Union Wage Policy*, Berkeley; Los Angeles 1948.

Frederick Zeuthen, *Problems of Monopoly and Economic Warfare*, London 1930.

Firestone Tire and Rubber Company, *67th Annual Report*, Akron, Ohio 1967.

U.S. Bureau of Labor Statistics, *Wage Chronology, Firestone Tire and Rubber Company and B. F. Goodrich Company (Akron Plants) 1937-73*, Washington 1972.

The Effects of a Firm's Investment and Financing Decisions on the Welfare of its Security Holders

By EUGENE F. FAMA*

In their classic article, Franco Modigliani and Merton H. Miller showed that in a perfect capital market, and given some other peripheral assumptions, the financing decisions of a firm are of no consequence. Substantial controversy followed, centered in large part on which of the peripheral assumptions are important to the validity of the theorem. For example, Joseph Stiglitz (1969, 1974) argues that in addition to a perfect market, the critical assumption is that bonds issued by individuals and firms are free of default risk. However, in chapter 4 of our book, Miller and I show that the theorem holds when debt is risky as long as stockholders and bondholders protect themselves from one another with what Fama and Miller (hereafter noted F-M) call "me-first rules."

This paper shows that me-first rules are also unnecessary. Propositions about the irrelevance of the financing decisions of firms can be built either on the assumption that investors and firms have equal access to the capital market or on the assumption that no firm issues securities for which there are not perfect substitutes from other firms. With either approach one can show that if the capital market is perfect, then (a) a firm's financing decisions have no effect on its market value, and (b) its financing decisions are of no consequence to its security holders.

The paper begins with a review of existing capital structure theorems, focusing on the work of Stiglitz and F-M. The discussion of old results has two purposes. The literature in this area has tended to become increasingly mathematical. One of the goals here is to show that the capital structure propositions in fact rest on simple economic arguments. Examining previous results also helps put the new results to be presented into perspective.

Finally, F-M and Stiglitz (1972) note that when firms can issue risky debt, the market value rule for the investment decisions of firms is ambiguous. With risky debt, maximizing stockholder wealth, bondholder wealth, or the combined wealth of bondholders and stockholders can imply three different investment decisions. Stiglitz argues that firms are likely to maximize stockholder wealth, even though this might be less economically efficient than maximizing combined stockholder and bondholder wealth. Miller and I leave the issue unresolved. I argue here that maximizing combined stockholder and bondholder wealth is the only market value rule consistent with a stable equilibrium, and that in its capacity as price setter the market can provide incentives for firms to choose this rule.

I. Arbitrage Proofs of the Market Value Proposition

Much of the early literature is concerned with the proposition that the market value of a firm is unaffected by its financing decisions, and most of the early proofs use arbitrage arguments. The general idea is that if the financing decisions of a firm affect its market value, there are arbitrage opportunities that can be used to produce costless instantaneous increases in wealth. Since the existence of such opportunities is inconsistent with equilibrium in a perfect

*Graduate School of Business, University of Chicago. This research is supported by the National Science Foundation. I am grateful for the comments of R. Ball, M. Blume, G. Borts, H. DeAngelo, N. Gonedes, R. Hamada, M. Jensen, S. Ross, M. Scholes, G. W. Schwert, and R. Weil. If I have any clear thoughts on the subject matter of this paper, they are due in large part to discussions with Merton H. Miller.

capital market, one can conclude that the market value of a firm is unaffected by its financing decisions. Examples of this approach are the original "risk class" model of Modigliani and Miller and the "states of the world" model of Jack Hirshleifer (1965, 1966).

In all of the arbitrage proofs of the market value proposition, there are five common assumptions:

Assumption 1: *Perfect Capital Market.* There are no transactions costs to investors and firms when they issue or trade securities; bankruptcy likewise involves no costs; there are no taxes; and there are no costs in keeping a firm's management to the decision rules set by its security holders. The perfect capital market assumption is maintained throughout the paper. Thus, I shall not discuss the interesting problems that arise from the differential treatment of corporate dividend and interest payments in computing corporate taxes, or the problems that arise from the differential treatment of dividends and capital gains in computing personal taxes. Nor shall I discuss any effects of bankruptcy costs or managerial agency costs on the nature of optimal investment and financing decisions by firms.

Assumption 2: *Equal Access.* Individuals and firms have equal access to the capital market. This means that the types of securities that can be issued by firms can be issued by investors on personal account. For example, suppose an investor owns the same proportion of each of a firm's securities, so that he has a direct share in the firm's activities. Equal access implies that, using the firm's securities as exclusive collateral, the investor can issue the same sort of securities as the firm. If firms can issue securities that contain limited liability provisions, such provisions can also be included in securities issued by investors against their holdings in firms. Moreover, the prices of securities are determined by the characteristics of their payoff streams and not by whether they are issued by investors or firms. Equal access could logically be included as a characteristic of a perfect capital market, but it plays such an important role in capital structure propositions that it is stated separately.

Assumption 3: *Complete Agreement or Homogeneous Expectations.* Any information available is costlessly available to all market agents (investors and firms), and all agents correctly assess the implications of the information for the future prospects of firms and securities. For most of what we do, it would be sufficient to assume that all market agents can correctly determine when securities issued by different investors and firms are perfect substitutes, but it seems at best a short step from this to complete agreement. A perfect capital market could be taken to imply complete agreement, but it is common in the literature to state the two as separate assumptions.

Assumption 4: *Only Wealth Counts.* Aside from effects on security holder wealth, the financing decisions of a firm do not affect the characteristics of the portfolio opportunities available to investors. Thus the effects of a firm's financing decisions on the welfare of its security holders can be equated with effects on security holder wealth. This assumption is only precise in the context of models that say which characteristics of portfolio opportunities are of concern to investors. We need not be so specific. For our purposes it is sufficient to assume that the capital market satisfies whatever conditions are necessary to ensure the desired correspondence between wealth and welfare. Moreover, we shall see that one of the contributions of more recent treatments of capital structure propositions is to show that this assumption is unnecessary.

Assumption 5. *Given Investment Strategies.* To focus on the effects of a firm's financing decisions on the welfare of its security holders, all proofs of capital structure propositions take the investment strategies of firms as given. Although decisions to be made in the future are unknown, the rules

that firms use to make current and future investment decisions are given. In addition, investment decisions are made independently of how the decisions are financed. In the last section of the paper, we consider the nature of optimal investment strategies for firms.

Stiglitz (1974, Theorem 2) gives the most general arbitrage proof that Assumptions 1-5 imply that the market value of a firm is unaffected by its financing decisions. Suppose there is an optimal capital structure for the firm, but the firm does not choose this capital structure. Any investor can provide the optimal capital structure to the market by buying equal proportions of the firm's securities and then issuing the optimal proportions on personal account. If the market value of the firm were less than the value implied by an optimal capital structure, by providing the optimal capital structure to the market, the investor could earn an arbitrage profit. Since every investor has an incentive to exploit such opportunities and since exploitation is costless, their existence is inconsistent with a market equilibrium. In equilibrium, the market value of a firm is always the value implied by an optimal capital structure, irrespective of the capital structure chosen by the firm. Thus, at least with respect to its effects on the firm's market value, any choice of capital structure by the firm is as good as any other.

II. Market Value and Security Holder Indifference

In the fourth chapter of our book, Miller and I show that the absence of a relationship between a firm's market value and its financing decisions does not in itself imply that the financing decisions are of no consequence to the firm's security holders. When the firm can issue risky debt, it may be able to use its financing decisions to shift wealth from its bondholders to its stockholders or vice versa.

To illustrate, assume a discrete time world in which the firm can issue two general types of securities, bonds and common stock. Given a perfect capital market and a market where the financing decisions of a firm do not affect the important characteristics of the portfolio opportunities available to investors, there is nothing the firm can do with its financing decisions at time t that will help or hurt investors who buy the firm's securities at time t. Thus it suffices to examine the effects of the firm's financing decisions at t on the wealths of investors who have held its securities from t − 1.

Let $S_{t-1}(t)$ and $B_{t-1}(t)$ be the market values at time t of the firm's common stock and bonds outstanding from t − 1. The combined value of these old stocks and bonds at t is the market value of the firm $V(t)$, less the value of new bonds issued at t, $b(t)$, less the market value of new common stock $s(t)$:

$$(1)\quad S_{t-1}(t) + B_{t-1}(t) = V(t) - b(t) - s(t)$$

The firm also makes dividend and interest payments at t, and we assume these are made only on securities outstanding from t − 1. Total dividend payments $D(t)$ and interest payments $R(t)$ are defined by

$$(2)\quad D(t) + R(t) = X(t) - I(t) + b(t) + s(t)$$

where $X(t)$ is net cash income at t (cash revenues minus cash costs), and $I(t)$ is the cash outlay for investment. Adding (1) and (2), the total wealth at time t associated with common stock and bonds outstanding from t − 1 is

$$(3)\quad [D(t) + S_{t-1}(t)] + [R(t) + B_{t-1}(t)] = X(t) - I(t) + V(t)$$

Since all capital structure propositions take the firm's investment strategy as given, $I(t)$ does not depend on financing decisions at t. The net cash earnings $X(t)$ are the result of past investment decisions and so are independent of financing decisions at t. Assumptions 1-5 ensure that the value of the firm $V(t)$ is unaffected by its financing decisions. Since $X(t)$, $I(t)$, and $V(t)$ are all independent of financing decisions at t, we can conclude from (3) that the combined

wealth of old bondholders and stockholders at time t is independent of the firm's financing decisions at t.

However, there might be financing decisions that the firm can make at time t that change the nature of the claims represented by the bonds outstanding from t – 1 and so shift wealth from bondholders to stockholders or vice versa. For example, suppose the firm's old bonds are free of default risk if no new debt is issued, but the firm can issue new debt that has the effect of imposing default risk on the old bonds. The new debt thus brings about a change in the characteristics of the old debt which we would expect to lead to a lower value of $B_{t-1}(t)$. Since the combined wealth of the old bonds and stocks is independent of the financing decision, issuing the new debt has the effect of shifting wealth from the old bondholders to the old stockholders. Alternatively, suppose the old debt is already subject to default risk, and at time t the firm retires some of it but not the entire amount. In the event of bankruptcy at a future date, each of the remaining bonds recovers more than if some of the old bonds are not retired at t. When a firm announces such a financing decision at t, we would expect the value $B_{t-1}(t)$ of all the old bonds to be higher than when no retirement takes place. Thus given constant total wealth, the financing decision implies a shift of wealth from the old stockholders to the old bondholders. In short, the fact that the market value of a firm is independent of its financing decisions does not necessarily imply that the financing decisions are a matter of indifference to the firm's security holders.

Given the world of Assumptions 1–5, the indifference proposition will hold if we restrict the types of securities that can be issued by firms so as to guarantee that the characteristics of the payoffs on the firm's old bonds are unaffected by its financing decisions at t. One way to accomplish this is to assume that all debt is free of default risk, which is the approach taken by Stiglitz (1969, 1974). In chapter 4 of our book, however, Miller and I show that the desired result is obtained when investors protect themselves with me-first rules. For example, bondholders insist that any new debt issued is junior to existing debt—in the event of bankruptcy, older bonds are paid off before newer bonds. The stockholders in their turn insist that the firm does not use its financing decisions to improve the positions of any bondholders. For example, if the firm wants to retire debt before its maturity, junior issues must be retired before senior issues, and any issues retired must be retired in full. We formalize these statements with a new assumption.

Assumption 6: A firm's stockholders and bondholders protect themselves from one another with costlessly enforced me-first rules which ensure that the characteristics of the payoffs on the firm's outstanding bonds are unaffected by changes in its capital structure.

In sum, Assumptions 1–5 are sufficient to conclude that the market value of a firm is unaffected by its financing decisions. Risk-free debt or the me-first rules of Assumption 6 then lead to the somewhat stronger conclusion that the financing decisions of the firm are a matter of indifference to all of its security holders.

III. The Irrelevance of a Firm's Dividend Decisions

A firm's dividend decision at any time t is part of its financing decision. The preceding analysis implies that when a firm's securities are protected by me-first rules, the firm's dividend decision at t determines how the wealth of its shareholders is split between $D(t)$ and $S_{t-1}(t)$, but the sum of the two components of shareholder wealth is unaffected by the dividend decision. In short, dividend decisions are a matter of indifference to the firm's security holders whenever financing decisions are a matter of indifference.

However, dividend decisions can be a matter of indifference even when other aspects of the firm's financing decisions are of some consequence. Consider a world

where the market value of a firm $V(t)$ is unaffected by its financing decisions, but the firm has risky debt outstanding which is not protected by me-first rules. By issuing more or less new bonds $b(t)$ at time t, the firm can affect the value of its old bonds $B_{t-1}(t)$, which in turn affects the split of wealth between its old bonds and its old stock. Any such effects on the wealths of old bonds and stocks are, however, due entirely to the choice of $b(t)$. Since the firm can issue more or less new stock $s(t)$ at time t, we can see from equation (2) that the choice of $b(t)$ need not affect the decision about the dividend $D(t)$. We can see from equations (1) to (3) that given any decision about $b(t)$ and its implication for $B_{t-1}(t)$, the dividend decision again just affects the split of shareholder wealth between dividends and capital value.

Keep in mind that we are taking the investment strategy of the firm as given. For example, if a firm that has risky bonds outstanding unexpectedly increases its dividend by selling off assets, there is a shift in wealth from bondholders to stockholders. However, the shift should be attributed to the investment decision, the sale of assets, rather than to the dividend decision since the same shift of wealth takes place, but in the form of a capital gain instead of a dividend, if the firm announces that the proceeds from the sale of assets will be used to repurchase shares.

IV. Dropping the "Only Wealth Counts" Assumption

Beginning with Modigliani and Miller, proofs of capital structure propositions generally include the Assumption 4 that aside from effects on security holder wealth, the financing decisions of firms do not affect the characteristics of the portfolio opportunities available to investors. Thus, the effects of financing decisions on security holder welfare can be evaluated in terms of their effects on security holder wealth. An exception to this approach is Stiglitz (1969, 1974) who shows that assumptions that lead to capital structure propositions also imply a world where the portfolio opportunities facing investors are unaffected by the financing decisions of firms. Formally:

THEOREM 1: *Suppose the capital market is perfect in the sense of Assumption 1, the equal access and complete agreement provisions, Assumptions 2 and 3, hold, the investment strategies of firms are given in the sense of Assumption 5, and debt is either free of default risk or investors insist on the me-first rules of Assumption 6. Then the characteristics of a general equilibrium, that is, the market values of firms, the positions that investors take in firms and the costs of these positions, are unaffected by the financing decisions of firms. Thus, the financing decisions of firms are of no consequence to investors.*

The intuition of the argument of Stiglitz' theorem is that when investors and firms have equal access to the capital market, the positions in firms that can be created and traded among investors are determined by the investment strategies of firms, and the possibilities are the same for any set of financing decisions by firms. Thus, the financing decisions of firms have no effect on the set of general equilibria that can be achieved in the capital market.

Moreover, once a general equilibrium has been achieved, implying an optimal set of holdings in firms by investors, there is no reason why changes in the financing decisions of firms should move the market to a different general equilibrium. When firms perturb a general equilibrium by changing their financing decisions, their actions neither expand nor contract the types of positions in firms that can be created by investors. It follows that an optimal response to the changes in the financing decisions of firms occurs when the general equilibrium remains unchanged. Specifically, the market responds by leaving the values of firms and their previously existing bonds unchanged. Investors respond by exactly reversing the changes in the financing decisions of firms on personal account so that the positions of investors in firms are unaffected by the changes in the financing decisions of firms.

The formal proof of Theorem 1 requires that changes in the financing decisions of firms can be reversed by investors on personal account. For this, the equal access Assumption 2 is required, but it is also assumed either that bonds are free of default risk (the assumption that Stiglitz (1974) uses in his proof of Theorem 1) or that investors insist on and costlessly enforce appropriate me-first rules (the extension of Stiglitz's analysis suggested by F-M). In the presence of risky bonds and in the absence of me-first rules, the firm can use changes in its financing decisions to, in effect, expropriate the positions of bondholders to the benefit of stockholders, or vice versa. And the expropriations cannot always be neutralized by investors on personal account.

For example, suppose the firm increases the dividend paid to stockholders at time t by issuing new bonds that have the same priority as the firm's old bonds in the event of bankruptcy. Even if the shareholders use the increase in dividends to repurchase the new bonds issued by the firm, things are not as they were. The new bonds are still outstanding, so that in the event of bankruptcy each of the old bonds gets less than if no new bonds are issued. By issuing new bonds that have equal priority with the old bonds, the firm has expropriated part of the holdings of the old bondholders to the benefit of its stockholders. Other examples, some involving expropriations of stockholder positions to the benefit of bondholders, are easily constructed.

V. Capital Structure Propositions without Me-First Rules

The assumptions that debt is free of default risk or security holders protect themselves with me-first rules are, however, arbitrary restrictions on the types of securities that can be issued. Some firms or investors may want to issue unprotected bonds, and, appropriately priced, other investors may be willing to hold them. It is now argued that such restrictions on investment opportunities are unnecessary, and this is the first new result of the paper.

THEOREM 2: *Suppose the capital market is perfect in the sense of Assumption 1, the equal access and complete agreement provisions, Assumptions 2 and 3, hold, and the investment strategies of firms are given in the sense of Assumption 5. Then the characteristics of a general equilibrium, that is, the market values of firms, the positions that investors take in firms and the costs of these positions, are unaffected by the financing decisions of firms. Thus, the financing decisions of firms are of no consequence to investors.*

To establish the theorem we return to time 0, the time when the first firms are organized and before they have issued any securities. The firms choose their investment strategies and then they go into the capital market for the resources to finance these investment strategies. At this point it is clear that given a perfect capital market and given equal access to the market by individuals and firms, the financing decisions of firms have no effect on the nature of a general equilibrium. The positions in firms that investors create and hold, the prices of these positions, and thus the market values of firms are independent of the financing decisions of firms

If unprotected securities are issued at time 0, then when time 1 comes along firms may be able to use their financing decisions to affect the positions of their security holders. When they hold the securities of a firm that are not protected by me-first rules, investors would of course prefer that the firm not engage in financing decisions at time 1 that have the effect of expropriating their positions; or, they would rather that the firm expropriate to their benefit the positions of other investors. But all of this is irrelevant, once we reconsider how it happened that at time 0 some investors put themselves into positions that could be expropriated at time 1. In an equal access market, the financing decisions of firms affect neither the variety of securities that could be traded at time 0 nor the instruments that are chosen by investors. If the positions that investors want to hold in firms are not offered by the firms, investors

can buy up the securities of firms and create their desired positions in trades among themselves. Thus, the positions, protected and unprotected, that investors take in firms at time 0 are the same irrespective of the financing decisions of firms at time 0. If at time 1 some investors profit from or are hurt by unprotected positions taken at time 0, all of this happens to exactly the same extent for any set of financing decisions by firms at time 0.

Likewise, at time 1 firms cannot use their financing decisions to affect the positions in firms that investors choose to carry forward to time 2. Given an equal access market, investors can refinance any firm, buying equal proportions of all its securities, and then issuing preferred proportions on personal account Thus the types and quantities of claims against firms that investors carry forward from time 1 to time 2 are independent of the financing decisions of firms at time 1. If expropriations take place at time 2 as a result of positions taken at time 1, the same investors are helped or hurt by these expropriations and to exactly the same extent when the unprotected securities are issued at time 1 by firms as when they are issued by investors in trades among themselves.

The arguments are general When investors and firms have equal access to the capital market, at any point in time the positions that investors take in firms, the prices of these positions and thus the market values of firms are unaffected by the financing decisions of firms. Since the financial history of any investor what happens to him in the market through time is unaffected by the financing decisions of firms, the financing decisions of firms are of no consequence to investors.

In all versions of the capital structure propositions discussed so far, equal access to the capital market by investors and firms is assumed. However, the assumption is stronger when debt is neither free of default risk nor protected by me-first rules. One is likewise leaning harder on the complete agreement assumption. Investors must be able to specify the details of potentially expropriative contracts in the same way as firms. If investors issue unprotected bonds against their holdings in firms, they subsequently expropriate (for example, issue more unprotected bonds) in the same circumstances as would the firms. This requires either that the conditions or states of the world in which expropriations will take place at any time t are stated explicitly in loan contracts or that investors make accurate assessments of the probabilities and extent of expropriations in different future states of the world. Probabilistically speaking, neither issuers nor purchasers of loan contracts are ever "fooled" by anything that happens during the life of a contract, and the price of a contract always properly reflects the possibilities for future expropriations.

I now show that the capital structure propositions can be established without the equal access assumption. The cost, however, is a new assumption which precludes a firm from issuing any securities monopolistically. In effect, we set up conditions that lead to a capital market which is perfectly competitive with respect to the financing decisions of a firm.

VI. Capital Structure Propositions without Equal Access

In Theorem 2, as in Theorem 1, the portfolio opportunities facing investors turn out to be independent of the financing decisions of firms. However, firms can still be monopolists in their investment decisions. A firm may have access to investment opportunities that allow it to create securities with payoff streams whose characteristics cannot be replicated by other firms. Nevertheless, when there is equal access to financial markets, investors can issue the same claims against their holdings in firms that the firms themselves can issue. As a consequence, once firms have chosen their investment strategies, there is nothing further they can do through their financing decisions to affect the opportunity set facing investors.

If this result is to hold when the equal ac-

cess assumption is dropped, we must restructure the world in such a way that the actions that investors (with equal access) take to free the investment opportunity set from any effects of financing decisions by firms, can be taken instead by firms. To accomplish this, firms are no longer allowed to issue securities for which there are not perfect substitutes issued by other firms. This implies that firms can no longer have monopolistic access to investment opportunities. Firms must also be given the motivation to act in the manner that leads to the validity of the capital structure propositions. In contrast, in an equal access world, once firms choose their investment strategies, what then happens when they get themselves to the capital market is beyond their control.

The specific new assumptions are:

Assumption 7: No firm produces any security monopolistically. There are always perfect substitutes issued by other firms. Moreover, if a firm shifts its capital structure, substituting some types of securities for others, its actions can be exactly offset by other firms who carry out the reverse shift, with the result that aggregate quantities of each type of security are unchanged.

Assumption 8: The goal of a firm in its financing decisions is to maximize its total market value at whatever prices for securities it sees in the market. Since firms are shown to be perfectly competitive in the capital market, the assumption is unobjectionable.

The arguments in the proof of the theorem that follows are similar to those used by the author and Arthur Laffer in discussing sufficient conditions for perfect competition in product markets in a world of perfect certainty. Also relevant are papers by the author (1972) and Fischer Black and Myron Scholes.

THEOREM 3: *Suppose the capital market is perfect in the sense of Assumption 1, the complete agreement assumption, Assumption 3, holds, the investment strategies of firms are given in the sense of Assumption 5, and Assumptions 7 and 8 also hold. Then given a general equilibrium in the capital market at any time* t: (a) *The market value of a firm is unaffected by changes in its financing decisions;* (b) *the financing decisions of a firm are of no consequence to investors; that is, the firm's financing decisions do not affect what happens to any investor through time;* and (c) *the capital market is perfectly competitive in the sense that aggregate supplies and prices of different types of securities are unaffected by changes in the financing decisions of a firm.*

Consider first the case where debt is free of default risk or investors protect themselves from one another with the me-first rules of Assumption 6. Suppose the capital market achieves a general equilibrium at time t and then, for whatever reason, some firm perturbs the equilibrium by changing its capital structure.

In the original equilibrium, firms, including the firm that subsequently shifts, chose securities so as to maximize their market values at the original equilibrium values of security prices. This means that at the original prices, the new securities that the shifting firm issues had exactly the same market value as the securities it no longer issues. It also means that the market can achieve a "new" general equilibrium if other firms instantly respond to the disturbance of the initial equilibrium by exactly offsetting the change in the shifting firm's capital structure, and if the prices of securities remain at their old equilibrium values. When this happens, the market value of any firm is the same as it was in the old general equilibrium, and firms have no further incentives to change their capital structures. In addition, since debt is assumed to be either free of default risk or securities are protected by me-first rules, the wealths of individual investors are the same in the new general equilibrium as in the old. Since the aggregate supplies and prices of different securities are unchanged, each investor can choose a portfolio identical to the one

chosen in the initial general equilibrium; just the names of the firms issuing particular types of securities may be different. In short, with me-first rules and the perfectly competitive capital market produced by the offsetting financing decisions of other firms, investors are completely immunized from any effects of shifts in the financing decisions of any firm.

The same analysis applies in the absence of me-first rules, once we understand the restrictions implied by the perfect substitutes Assumption 7. In particular, the fact that different firms issue securities at time t − 1 that are perfect substitutes does not imply that these firms make the same financing decisions at time t. However, if unprotected securities issued by different firms at t − 1 are perfect substitutes, any expropriations that take place at time t must be the same for all of these firms. It follows that if a firm issues unprotected securities at any time t − 1, the expropriations that take place in any given state of the world at time t must be the same for all financing decisions that the firm might make in that state at t.

Suppose now that time t comes along, the state of the world is known, firms make their financing decisions, and a general equilibrium set of securities prices and values of firms is determined. Some firm then perturbs the general equilibrium by shifting its capital structure. Given what was said above, even though the firm may have unprotected securities in its capital structure, the shift cannot cause expropriations of security holder positions beyond those associated with the firm's original financing decisions at t. Thus, just as in the case where debt is risk free or securities are protected by me-first rules, the market can reattain a general equilibrium if other firms exactly offset the change in the shifting firm's capital structure, leaving aggregate supplies and prices of different securities unchanged. Since no new expropriations take place, the wealths of investors are also unchanged, and each investor can choose a portfolio identical to the one chosen in the initial general equilibrium.

In the initial general equilibrium that follows the occurrence of a state of the world at time t, the positions of a firm's security holders are, of course, affected by any expropriative financing decisions. But in the world of the complete agreement Assumption 3, investors properly assessed the possibilities for future expropriations when they decided to hold the firm's securities at time t − 1, and these possibilities were properly reflected in the prices of the securities at t − 1. If the firm hadn't issued these potentially expropriative securities, its security holders would have purchased perfect substitutes from other firms. Thus the financing decisions of any firm are of no consequence to any investor in the sense that what happens to any investor through time happens irrespective of the financing decisions of any particular firm.

VII. Some Perspective on Capital Structure Propositions

Given a perfect capital market, and given the investment strategies of firms, there are two approaches that lead to the conclusions that the market value of a firm is unaffected by its financing decisions, and a firm's financing decisions are of no consequence to its security holders. One approach is based on the assumption that investors and firms have equal access to the capital market. The other assumes that no firm offers securities to the market for which there are not perfect substitutes from other firms. The fundamental argument in both approaches is that, given the investment strategies of firms, there are mechanisms that insulate the opportunity set facing investors from any effects of the financing decisions of firms. With the equal access assumption, the offsetting actions that produce this result can come from investors or firms, while in the perfect substitutes approach, changes in the financing decisions of a firm are offset by other firms.

The types of capital structure propositions obtained with the two approaches are somewhat different. With equal access one gets statements about the effects of the

financing decisions of all firms. When investors and firms have equal access to the capital market, then given the investment strategies of firms, the positions in firms that can be traded among investors are independent of the financing decisions of firms. As a consequence, the characteristics of a general equilibrium in the capital market are unaffected by the financing decisions of firms. In contrast, with the perfect substitutes approach, only firms issue securities so one can't conclude that the characteristics of a general equilibrium are independent of the financing decisions of all firms. One is limited to partial equilibrium statements about the irrelevance of the financing decisions of any individual firm.

The analysis here goes beyond earlier treatments in several respects. First, although earlier approaches generally use both assumptions in one form or another, it is evident from the work of Stiglitz (1969, 1974) that in an equal access market, the validity of the capital structure propositions does not also require the perfect substitutes assumption. However, Stiglitz argues that it is necessary to assume debt is risk free if the financing decisions of firms are to be a matter of indifference to security holders. The analysis here shows that in an equal access market, even the me-first rules of F-M are unnecessary restrictions on the types of securities that can be issued. In essence, in an equal access market investors can and will choose the same positions, protected and unprotected, irrespective of the financing decisions of firms. Thus, the fact that firms might issue unprotected securities does not invalidate the proposition that the financing decisions of firms are a matter of indifference to investors.

In a recent paper, Frank Milne argues that with the perfect substitutes assumption, the proposition that the market value of a firm is independent of its financing decisions does not also require the equal access assumption. However, Milne's framework is less general than that examined here. First, he allows unrestricted short selling of all securities, an assumption close to equal access. To emphasize the power of the perfect substitutes assumption, in the analysis presented here securities are only issued by firms. Second, Milne assumes that the capital market is perfectly competitive, whereas we show how the actions of firms lead to a world where the total supplies and prices of securities of different types are unaffected by the financing decisions of any individual firm. Showing how the existence of perfect substitutes leads to such a strong form of perfect competition seems a substantial enrichment of the analysis. Finally, Milne works in a one-period context and investors do not come into the period already holding the securities of firms. In this world, the analytical difficulties that arise from potential expropriations of security holder positions never have to be faced. In contrast, I analyze the capital structure propositions in a multiperiod framework where firms are allowed to issue unprotected securities. It is shown that with the strong form of perfect competition in the capital market that arises from the perfect substitutes assumption, the financial history of any investor, that is, the protected and unprotected portfolio positions that he takes through time, are unaffected by the financing decisions of any individual firm.

Many have quarreled with the realism of the equal access assumption. (See, for example, the comments of David Durand on the Modigliani and Miller paper.) One can certainly also quarrel with the perfect substitutes assumption. It would seem that if for any securities issued by a firm there are perfect substitutes issued by other firms, then either there exist risk classes of firms in the sense of Modigliani and Miller (that is, there are classes of firms wherein the net cash flows of different firms are perfectly correlated) or the markets for contingent claims discussed by Hirshleifer cover all possible future states of the world. The existence of such risk classes or of complete markets for contingent claims is questionable.

In economics, however, formal propositions never provide pictures of the world that are realistic in all their details. The role of such propositions is to pinpoint the fac-

tors that can lead to certain kinds of results. In this view, the analysis of capital structure propositions suggests two factors that push the capital market toward equilibria where the market values of firms are independent of their financing decisions, and where the financing decisions of firms are of no consequence to their security holders. The first factor covers any possibilities investors have to issue claims against the securities of firms that they hold. The second is the natural incentive of firms to provide the types of securities desired by investors, and the ability of firms to provide securities that are close substitutes for those of other firms. In pure form, and in combination with a perfect capital market where contracts are costlessly written and enforced, either of these factors leads to irrelevance of capital structure propositions. In less pure form, but perhaps acting together, they are factors that help to push the market in the direction of the capital structure propositions.

VIII. The Market Value Rule for Investment Decisions

The previous sections discuss the financing decisions of firms, given their investment strategies. I turn now to problems that arise in determining an optimal investment strategy when the capital market is perfect and when a firm can affect the portfolio opportunities facing its security holders only through the effects its investment decisions have on the wealths of its security holders. All other characteristics of portfolio opportunities are assumed to be unaffected by the investment and financing decisions of the firm.

Given that the investment decisions of a firm only affect the wealths of its security holders, the objectives of the security holders are clear. More wealth is better than less. In chapter 4 of our book, however, Miller and I point out that the "maximize securityholder wealth" rule can be ambiguous when the firm has risky debt. The firm might be able to use its investment decisions to make its previously issued bonds more or less risky and so to shift wealth from bondholders to stockholders or vice versa. One can easily construct examples where the rules "maximize stockholder wealth," "maximize bondholder wealth," and "maximize combined stockholder-bondholder wealth" all lead to different investment decisions.

A. *The Pressure of Possible Takeovers*

We can apply the argument of Ronald Coase to show that of the three market value rules, only the rule maximize combined stockholder-bondholder wealth is consistent with a stable capital market equilibrium. Note first that when the capital market is perfect and when the characteristics of portfolio opportunities are independent of the actions of any individual firm, there is nothing the firm can do with its investment decision at t to help or hurt investors who buy its securities at t. Thus it suffices to examine the effects of the firm's investment decision at t on investors who have held its securities from t − 1.

From equation (3), the combined wealth at time t of the firm's bonds and stocks outstanding from t − 1 is $X(t) + V(t) - I(t)$. Since net cash earnings $X(t)$ are assumed to result from past decisions, they are unaffected by the investment decision at t. Thus maximum combined stockholder-bondholder wealth implies maximizing $V(t) - I(t)$, the excess of the market value of the firm at t over the investment outlays needed to generate that market value.

Suppose the firm is controlled by its stockholders, and they choose the rule maximize stockholder wealth. It will pay for the firm's bondholders to buy out the stockholders, paying them the value their shares would have under the rule maximize stockholder wealth. If the bondholders then maximize $V(t) - I(t)$, we can see from (3) that their wealth is larger than if they had allowed the shareholders to proceed with the investment rule maximize stockholder wealth. The same arguments apply, but with the roles of the stockholders and bondholders reversed, when the firm is initially controlled by its bondholders who

wish to follow the rule maximize bondholder wealth. Alternatively, if the firm announces an investment rule other than "maximize $V(t) - I(t)$," it pays for outsiders to buy up the firm's securities and then to switch to the rule maximize $V(t) - I(t)$. The outsiders can even afford to pay a premium for the firm as long as it is no greater than the difference between the maximum value of $V(t) - I(t)$ and the value of $V(t) - I(t)$ under the investment policy chosen by the firm.

B. *The Pressures Applied by the Market in its Capacity as Price Setter*

Potential takeovers are not the only pressure pushing the firm toward the investment rule maximize $V(t) - I(t)$. In its role as price setter, the market has an additional way to motivate the firm to maximize the total wealth of its security holders.

Consider the firm's bondholders. When the firm issues bonds, the price of a given promised stream of payments depends on the investment strategy that the market perceives the firm to follow. If the firm in fact follows this strategy, the investment strategy is of no consequence to the bondholders. If they had the choice again, with the same uncertainties about the future, they would choose to hold the firm's bonds or perfect substitutes for them. In a capital market where the investment and financing decisions of a firm do not affect the portfolio opportunities facing investors, such perfect substitutes exist or they can be created from the securities of other firms. Since the market for shares is likewise a market of perfect substitutes, given that a firm sticks to the investment strategy that investors perceive it to follow, the choice of strategy is of no consequence to those who purchase its shares when it is an ongoing firm. In this situation, the choice of an investment strategy by the firm affects only the firm's original shareholders or organizers, those who own the rights to its investment opportunities before any securities are issued.

Let us return, then, to the point, call it time 0, when the firm is organized. The firm wishes to choose the investment strategy that maximizes the wealth of its organizers. The wealth of the organizers is $V(0) - I(0)$, the difference between the value of the firm and the investment outlays necessary at time 0 to generate that value. Thus the optimal investment decision at time 0 is to maximize $V(0) - I(0)$.

The value of the firm $V(0)$ depends also on the investment strategy the market thinks the firm will follow at time 1. Since the wealth at time 1 of securities outstanding from time 0 is $X(1) + V(1) - I(1)$, the value of the firm at time 0 is just the market value at time 0 of the distribution of $X(1) + V(1) - I(1)$. The earnings $X(1)$ observed at time 1 are a consequence of the investment decision taken at time 0. In every possible state of the world at time 1, the policy "maximize $V(1) - I(1)$" obviously produces as large a value of $V(1) - I(1)$ as any other investment strategy. It follows that if the firm's statements about investment policy are accepted by the market, the announcement at time 0 that the firm will maximize $V(1) - I(1)$ at time 1 maximizes the contribution of the investment decision at time 1 to $V(0)$ and thus to $V(0) - I(0)$.

Since the market value of the firm at time 1 is just the market value of the distribution of $X(2) + V(2) - I(2)$, $V(1)$ and thus $V(1) - I(1)$ depend in turn on the investment strategy that will be followed at time 2. Arguments analogous to those above imply that the announcement, at time 0, that the firm will maximize $V(1) - I(1)$ at time 1 implies the announcement, at time 0, that it will maximize $V(2) - I(2)$ at time 2. In short, to maximize $V(0) - I(0)$, the wealth of its organizers at time 0, the firm must convince the market that its investment strategy in each future period will likewise be maximize $V(t) - I(t)$. If the firm sticks to this strategy, this means that at any time t it chooses the investment decisions that maximize the combined wealth of bonds and stocks outstanding from $t - 1$.

Using the analysis of "agency costs" provided by Michael Jensen and William

Meckling one can argue that the essence of the potential problems surrounding conflicting stockholder-bondholder interests is that once time 0 passes it will be difficult for the stockholders to resist the temptation to try to carry out an unexpected shift from the rule maximize $V(t) - I(t)$ to the rule maximize stockholder wealth. But the market has the means to motivate firms to stay in line. To maximize $V(0) - I(0)$, the wealth of its organizers, the firm must convince the market that it will always follow the investment strategy maximize $V(t) - I(t)$. The market realizes that the firm might later try to shift to another strategy and it will take this into account in setting $V(0)$. To get the market to set $V(0)$ at the value appropriate to the strategy maximize $V(t) - I(t)$, the firm will have to find some way to guarantee that it will stay with this strategy

The important point is that the onus of providing this guarantee falls on the firm. In pricing a firm's securities, a well-functioning market will, on average, appropriately charge the firm in advance for future departures from currently declared decision rules. The firm can only avoid these discounts in the prices of its securities to the extent that it can provide concrete assurances of its forthrightness. Thus, firms have clear-cut incentives to evolve mechanisms to assure the market that statements of policy can be taken at face value, and they have incentives to provide these assurances at lowest possible cost. In a multiperiod world, this might not be so difficult since firms continually have opportunities to behave in ways that reinforce their credibility.

Remember also that if the firm does not follow the strategy maximize $V(t) - I(t)$, it pays for outsiders to acquire the firm and then switch to this strategy. The outsiders are then in the position of the firm's organizers. That is, the firm will not be priced at the value implied by the strategy maximize $V(t) - I(t)$ unless the market is convinced that the firm will adhere to this strategy in future periods. If other forms of assurance prove difficult or costly, one possibility is to finance the firm entirely with equity, or more generally, never to issue risky debt. Then the rules maximize stockholder wealth and maximize $V(t) - I(t)$ coincide.

REFERENCES

F. Black and M. Scholes, "The Effect of Dividend Yield and Dividend Policy on Common Stock Prices and Returns," *J Finan. Econ* , May 1974, *1*, 1-22.

R. H. Coase, "The Problem of Social Cost," *J Law Econ.*, Oct. 1960, *3*, 1 44.

D. Durand, "The Cost of Capital in an Imperfect Market: A Reply to Modigliani and Miller," *Amer. Econ. Rev.*, June 1959, *49*, 639-55.

Eugene F. Fama, "Perfect Competition and Optimal Production Decisions under Uncertainty," *Bell J. Econ.*, Autumn 1972, *3*, 509-30.

——— **and A. B. Laffer**, "The Number of Firms and Competition," *Amer. Econ. Rev* , Sept. 1972, *62*, 670-74.

——— **and Merton H. Miller**, *The Theory of Finance*, New York 1972.

J. Hirshleifer, "Investment Decisions under Uncertainty: Choice Theoretic Approaches," *Quart. J Econ.*, Nov. 1965, *79*, 509-36.

———, "Investment Decisions under Uncertainty: Applications of the State-Preference Approach," *Quart. J. Econ.*, May 1966, *80*, 237-77.

M. C. Jensen and W. H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *J. Finan. Econ.*, Oct. 1976, *3*, 305 60.

F. Milne, "Choice over Asset Economics: Default Risk and Corporate Leverage," *J. Finan. Econ.*, June 1975, *2*, 165-85.

F. Modigliani and M. H. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261-97.

J. C. Stiglitz, "A Re-Examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784-93.

———, "Some Aspects of the Pure Theory of Corporate Finance: Bankruptcies and Take-overs," *Bell J. Econ.*, Autumn 1972, *3*, 458 82.

———, "On the Irrelevance of Corporate Financial Policy," *Amer. Econ. Rev.*, Dec. 1974, *64*, 851-66.

Welfare Evaluation and the Cost-of-Living Index in the Household Production Model

By ROBERT A. POLLAK*

The household production model provides a new framework for the theory of the household, and most applications have focused on its implications for market and nonmarket behavior. In this paper I examine the consequences of the "new home economics" for welfare analysis, and in particular for the cost-of-living index.

In the household production framework market "goods" are combined with time to produce "commodities." These commodities, rather than the market goods, are the arguments of the household's preference ordering; the demand for goods and time is a "derived demand," since goods are not desired for their own sake, but only as inputs into the production of commodities.[1] This paper is an analysis of the implications of the household production model for welfare evaluation, not a critique of the model. Hence, it accepts the fundamental distinction between goods and commodities, and assumes that commodities as well as goods are observable and measurable.[2] The distinction between technology and tastes follows unambiguously from that between goods and commodities.

In orthodox demand theory the household's preference ordering is defined over the "goods space" and welfare analysis is based on those preferences. The cost-of-living index is defined as the ratio of the minimum expenditures required to attain a particular indifference curve of this preference ordering under two price regimes. In the household production model the preference ordering over the commodity space provides a corresponding basis for welfare evaluation. One way to extend the notion of the cost-of-living index to the new framework is to define it as the ratio of the minimum expenditures required to attain a particular indifference curve of this preference ordering under alternative price-technology regimes. This approach recognizes that the objects of the household's preferences are commodities rather than goods, and that both goods prices and the household's technology impose constraints on its choices; the resulting index, the "variable technology cost-of-living index," reflects changes in both the technical and market constraints facing the household.

For some purposes we are interested in separating the welfare impact of price changes from that of changes in the household's technology. For example, to measure the effectiveness of monetary and fiscal policy we want a "price index" which re-

*University of Pennsylvania This research was supported in part by the U.S Bureau of Labor Statistics and the National Science Foundation. I am grateful to Barbara Atrostic, Hugh Davies, W. E. Diewert, Stefano Fenoaltea, Franklin M Fisher, Robert P Inman, John Muellbauer, Stephen A. Ross, Paul J Taubman, Jack E. Triplett, the managing editor, and an anonymous referee. Their helpful comments and suggestions have substantially improved this paper, but I have not always followed their advice and none of them should be held responsible for the views expressed.

[1]The locus classicus of the household production literature is Gary Becker. In Kelvin Lancaster's model (1966a, b, 1971) goods possess "characteristics" which are often identified with Becker's commodities and the "technology" is linear. Becker often uses fixed coefficient production functions as an expositional device, but linear technology is not an integral part of his model. For a recent sympathetic statement of the household production approach, see Robert Michael and Becker. For a discussion of some of its limitations, see the author and Michael Wachter.

[2]The knowledge that production is going on within the household is of no help if our observations are limited to its purchases of market goods. The author and Wachter emphasize that the assumption that commodities are observable and measurable is crucial to exploiting the advantages of the household production model.

flects prices paid by the consumers but is independent of changes in household technology. On the other hand, if we use a household production framework to analyze the harm done by air pollution or the benefits of an outdoor recreation or child health project, we would want a measure which was independent of whatever price changes happen to have occurred. In all of these cases, we are concerned with "subindexes" which reflect some but not all of the variables which affect the household's welfare. Since measuring the welfare effect of price changes is the traditional concern of cost-of-living index theory, I shall emphasize the "constant technology cost-of-living index," an index which reflects only the effects of price changes, rather than the "constant price cost-of-living indexes" which reflect only technological and environmental changes. The constant technology index is defined as the ratio of the minimum expenditures required to attain a particular indifference curve under two price regimes when the household's technology is held fixed. This index, unlike the variable technology index, exhibits all of the properties familiar from traditional cost-of-living index theory.

Complete information about preferences is usually not available and the development of indexes which can be calculated with less information and are bounds on exact indexes is a major part of traditional index number theory. In the household production framework, complete information about preferences and technology is unlikely to be available, so there is a similar need for indexes which are bounds on exact indexes. I shall define a number of Laspeyres-type indexes in the household production framework and examine their relations to various exact indexes. I show that the "goods Laspeyres index"—defined in the obvious way—is an upper bound on the constant technology cost-of-living index. I define the "commodity Laspeyres index" in terms of the cost functions implied by the household's technology as the ratio of the cost of producing the reference commodity bundle under the comparison price-technology regime to its cost under the reference price-technology regime. This index is an upper bound on the variable technology cost-of-living index. If we not only know the household's technology but also have additional information about tastes—in particular, if an income-consumption curve in the commodity space is known—then the shadow prices of commodities can be used to construct an index which is a better bound on the variable technology cost-of-living index than that provided by the commodity Laspeyres index.

The literature on welfare evaluation in the household production framework is sparse. John Muellbauer provides one of few explicit discussions of the cost-of-living index in the household production model as a foundation for his analysis of quality and hedonic indexes. The other major treatment is that of Franklin Fisher and Karl Shell. The scope of their paper is somewhat broader than is indicated by its title—"Taste and Quality Change in the Pure Theory of the True Cost-of-Living Index" since they sometimes use the phrase "quality change" to refer to changes in the household's technology.[3] Muellbauer adopts this unfortunate terminology which serves to confuse two important but logically distinct problems. In my article (1975b) I suggest that quality and quality change should be thought of in terms of "variety choice," and discuss several formulations emphasizing "characteristics." The frequent identification of quality change with technical change may reflect the fact that both affect the welfare of households by altering the constraints which they face, yet neither is best thought of as a change in the price of a market good.[4] This paper originated in

[3]Fisher and Shell do not explicitly use the household production model, but it is clear from their examples (for example, refrigerators, ice cream, home-refrigerated ice cream) that they sometimes have in mind situations involving household production.

[4]In a model of variety choice, quality change may be treated as a change in the price of a market good, provided we treat the prices of unavailable goods as infinite. This formulation is of limited usefulness for empirically oriented work because most households consume only a small number of the varieties avail-

my dissatisfaction with the treatment of quality and quality change; specifically with attempts to justify it in terms of the household production framework. Although the analysis which follows originated as a prolegomenon to the discussion of quality and hedonic indexes, it applies to the entire range of phenomena which can be treated in the household production framework.

The organization of the paper is as follows: Section I summarizes the traditional theory of the cost-of-living index. Section II introduces the household production model and analyzes the variable and constant technology indexes. Section III considers the "goods cost-of-living index" and Section IV discusses bounds on the various exact indexes. Section V extends the discussion of bounds by showing that commodity shadow prices may sometimes be used to obtain better bounds than those established in Section IV. Section VI is a brief summary.

I. Traditional Theory

In this section I summarize the traditional theory of the cost-of-living index in its conventional setting: a model in which the goods the household purchases on the market are consumed directly rather than used as inputs in the production of commodities. Special attention is paid to the treatment of taste change, since it is instructive to contrast its treatment with that of technical progress in the household production framework.

We denote the goods vector by X, the household's preference ordering by R, and the corresponding utility function by $V(X)$.[5] The cost-of-living index is the ratio of the expenditures required to attain a particular indifference curve under two price regimes. We denote the cost of living index by

$$I(P^a, P^b, X^o, R) = \frac{E(P^a, X^o, R)}{E(P^b, X^o, R)}$$

where $E(P, X^o, R)$ is the "expenditure function," that is, the minimum expenditure required to attain the indifference curve of the goods vector X^o from the preference ordering R in price situation P.[6] The notation emphasizes that the index depends not only on the "comparison prices" P^a, and the "reference prices" P^b, but also on the choice of a base indifference map or preference ordering R, and the specification of a base indifference curve from that map.

The treatment of taste change in the cost-of-living index is often misunderstood, although careful expositions are presented in Fisher and Shell, and the author (1971). The cost-of-living index does not purport to measure or compare utilities or satisfaction actually attained in different periods or situations. Instead it compares two sets of constraints.[7] To do this it considers a particular preference ordering (R) and an indifference curve (X^o) from that ordering and asks what is the ratio of the minimum expenditure required to attain the indifference curve X^o of the preference ordering R in price situation P^a to that required in price situation P^b. If preferences are the same in both situations, it may seem "natural" to use the common preference ordering as the base for the cost-of-living index, although the theory does not require it.[8]

able. In my article (1975b) I discuss several "quality models" which avoid this difficulty by relying on characteristics.

[5]Assume that the household's preferences are well-behaved in the sense that they can be represented by a continuous utility function which is quasi concave and nondecreasing in its arguments.

[6]I have identified the base indifference curve by specifying a goods vector X^o which lies on it; replacing X^o by any other goods vector on the same indifference curve does not affect the expenditure function or the cost-of-living index. Instead of defining the expenditure function in terms of the weak preference relation R (at least as good as) we could equivalently begin with the utility function $V(X)$ and specify the base indifference curve by requiring $V(X) \geq V(X^o)$.

[7]For an alternative view based on comparing satisfaction actually attained in each situation and requiring the use of cardinal utility see Louis Phlips, ch. 11, and Phlips and Ricardo Sanz-Ferrer.

[8]This may be more transparent in the case of international than intertemporal comparisons. For example, to compare prices in Paris with those in Tokyo, one might use French tastes or Japanese tastes; but if the U.S. government wants to establish appropriate

If preferences are different, then the need to choose a base preference ordering cannot be ignored.

The Laspeyres index, $J(P^a, P^b, X^b)$, is defined by

$$J(P^a, P^b, X^b) = \frac{\Sigma x_k^b p_k^a}{\Sigma x_k^b p_k^b}$$

where X^b denotes the collection of goods bought in the reference situation.[9] It is well-known that the Laspeyres index is an upper bound on the cost-of-living index based on the reference period tastes R^b, evaluated at the reference period indifference curve X^b: $I(P^a, P^b, X^b, R^b)$.[10]

II. Variable and Constant Technology Indexes

In this section I introduce the problem of welfare evaluation in the household production framework and define the variable and constant technology cost-of-living indexes.

According to the household production view, the household purchases goods on the market and combines them with time in a household production process to produce commodities. These commodities, rather than the goods, are the arguments of the household's preference ordering, market goods and time are not desired for their own sake, but only as inputs into the production of commodities. I shall ignore the role of time in the household production process, and treat the production of commodities as depending only on market goods.[11] We denote the n market goods by $X = (x_1, \ldots, x_n)$, the m commodities by $Z = (z_1, \ldots, z_m)$, the household's preference ordering over commodity vectors by R and the corresponding utility function by $W(Z)$.[12] The household's technology is represented by a production set T. Thus, the "output-input" vector $(Z, X) \epsilon T$ if and only if the commodity collection Z is producible from the goods collection X. Unless explicitly stated to the contrary, constant returns to scale and/or the absence of joint production are *not* assumed, so there need not be a separate production function for each commodity.

The "cost function" $C(P, Z; T)$ is defined as the minimum cost of producing the commodity bundle Z with the technology T at goods prices P. The notation explicitly recognizes the role of the underlying technology. Formally,

DEFINITION: The *cost function* $C(P, Z; T)$ is defined by

$$C(P, Z; T) = \min \sum_{k=1}^{n} p_k x_k$$

subject to $(Z, X) \epsilon T$

Corresponding to the preference ordering R and the technology T we define the expenditure function $E(P, Z^o, R, T)$ which shows the minimum expenditure required to attain the indifference curve of Z^o.

salary differentials for its diplomats, the comparison should presumably be based on U.S. tastes. In intertemporal comparisons, the index may be based on any preference ordering, the choices are not restricted to reference and comparison period tastes.

[9]Throughout this paper I discuss the Laspeyres index and leave the Paasche index to the reader.

[10]The goods vector X^b plays a very different role in the Laspeyres index than it does in the cost-of-living index. In the cost-of-living index it specifies a particular indifference curve, and the replacement of X^b by another collection of goods on the same indifference curve does not alter the index. In the Laspeyres index it specifies a particular collection of goods, the collection actually bought in the reference situation. The fixed weight index constructed by replacing X^b by another collection of goods on the same indifference curve is not equal to the Laspeyres index and is not an upper bound on the cost-of-living index.

[11]I have ignored the role of time to avoid further complicating an already complicated notation. Time can be incorporated into the household production model in a straightforward way if we are willing to treat it as a "market good" whose price is the wage rate. Indeed, the allocation of time was the principal focus of Becker's seminal paper. The author and Wachter argue that joint production is likely to be pervasive when time plays a role in household production, and that joint production severely limits the usefulness of implicit "commodity prices" for demand analysis. In Section V, I argue that if the household's technology is known, commodity shadow prices can sometimes be used to obtain a useful bound on the cost-of-living index.

[12]The household's preferences over the commodity space are assumed to be well-behaved.

DEFINITION: The *expenditure function* $E(P, Z^o, R, T)$ is defined by

$$E(P, Z^o, R, T) = \min C(P, Z; T)$$

subject to ZRZ^o

If the household's technology remains unchanged, the obvious extension of the cost-of-living index to the household production model defines the index as the ratio of the expenditures required to attain a particular indifference curve under the two price regimes. But if the "comparison technology" T^α differs from the "reference technology" T^β, there are two plausible ways to formulate the index. One allows the index to reflect both changes in technology and changes in goods prices, while the other holds technology fixed and allows the index to reflect only changes in goods prices. Both are interesting and useful, but they answer different questions. If our purpose is to compare the expenditures required to attain a particular standard of living under alternative price-technology regimes, then the variable technology cost-of-living index is appropriate. If we want to compare the prices of consumer goods in two periods, the constant technology cost-of-living index is called for. Our first task is to set out these two indexes in a notation which makes it clear what the alternatives are.

DEFINITION: The *variable technology cost-of-living index* $I(P^a, P^b, Z^o, R, T^\alpha, T^\beta)$ is defined by

$$I(P^a, P^b, Z^o, R, T^\alpha, T^\beta) = \frac{E(P^a, Z^o, R, T^\alpha)}{E(P^b, Z^o, R, T^\beta)}$$

This index is the ratio of the minimum expenditure required to attain the indifference curve of Z^o at prices P^a with technology T^α, to the minimum expenditure required to attain the same indifference curve at prices P^b with technology T^β. Hence,

THEOREM 1: *Let $Z^{oa\alpha}(Z^{ob\beta})$ denote the minimum cost commodity bundle which attains the indifference curve of Z^o at prices $P^a(P^b)$ and technology $T^\alpha(T^\beta)$. Then*

$$I(P^a, P^b, Z^o, R, T^\alpha, T^\beta) = \frac{C(P^a, Z^{oa\alpha}; T^\alpha)}{C(P^b, Z^{ob\beta}; T^\beta)}$$

That is, the variable technology cost-of-living index can be written as the ratio of cost functions, provided they are evaluated at the "proper" points in the commodity space.

The variable technology cost-of-living index is not simply a "price index," for it depends not only on how P^a differs from P^b, but also on how T^α differs from T^β. It is tempting to identify the technology T^α with the price set P^a, and T^β with P^b; for reasons which will become clear below, I have not incorporated these restrictions into the definition of the variable technology index. The technologies associated with P^a and P^b are of special interest and are denoted by T^a and T^b, respectively.

The variable technology cost-of-living index $I(P^a, P^b, Z^o, R, T^a, T^b)$ yields a plausible answer to a version of the "cost-of-living" question in a household production model. It compares the reference constraint that corresponding to (P^b, T^b)— with the comparison constraint (P^a, T^a) by indicating the ratio of the expenditure needed to allow the household to attain the base indifference curve at prices P^a with technology T^a to that required at prices P^b with technology T^b. Not surprisingly, this index does not satisfy all the theorems of traditional index number theory.[13] For example, if all prices rise by 5 percent, then the index need not rise by 5 percent, and it might even fall; its behavior depends not only on how prices change, but also on the change in technology from T^b to T^a.

The constant technology cost-of-living index is a subindex of a complete cost-of-living index, namely, the variable technology cost-of-living index. In my article (1975a) I develop a theory of subindexes of the cost-of-living index; the discussion there is concerned with subindexes for subsets of goods (for example, clothing, footwear, men's shoes), but the analysis can be

[13]See the author (1971) for a detailed statement of these theorems.

applied when any of the variables which appear in the utility function or any of the constraints facing the household are not explicitly taken into account (for example, leisure, environmental variables or goods provided by the government without user charges). The complete index reflects changes in all of the constraints which the household faces, while subindexes are concerned only with changes in a subset of the constraints in the present case, with the market constraints but not with those imposed by the household's technology.[14]

The constant technology cost-of-living index is defined in terms of a particular technology (perhaps T^a, perhaps T^b, perhaps some other) and thus is independent of any actual changes in the household's technology. It is the ratio of the minimum expenditure required to attain the indifference curve of Z^o at prices P^a with technology T to that required at prices P^b with the same technology, T. Its most natural application is in comparing the reference constraint (P^b, T^b) with the hypothetical comparison constraint such as (P^a, T^b).

DEFINITION: The *constant technology cost-of-living index* $\bar{I}(P^a, P^b, Z^o, R, T)$ is defined by

$$\bar{I}(P^a, P^b, Z^o, R, T) =$$

$$I(P^a, P^b, Z^o, R, T, T) = \frac{E(P^a, Z^o, R, T)}{E(P^b, Z^o, R, T)}$$

This index, unlike the variable technology cost-of-living index, satisfies all of the theorems of cost-of-living index theory. The index depends on the choice of a "base technology," in much the same way that the traditional cost-of-living index depends on the choice of a base preference ordering, so it would not be correct to say that the index is independent of technology It is, however, independent of changes in the technology in the same sense as the traditional index is independent of changes in tastes; indeed, the treatment of technical change in the constant technology cost-of-living index is parallel to that of taste change in the traditional theory.

If the variable technology cost-of-living index is evaluated at a single technology $T^\alpha = T^\beta = T$, then it reduces to the constant technology index and satisfies all of the theorems of traditional cost-of-living index theory. If it is evaluated at a single set of prices $P^a = P^b = P$, then it reduces to the constant price cost-of-living index, an index which measures changes in the household's technology This index exhibits a number of properties which are desirable in an index of technical change For example, if the comparison technology is proportional to the reference technology (i.e., for all $\lambda > 0$, $(Z, \lambda X) \epsilon T^\alpha$ if and only if $(Z, X) \epsilon T^\beta$) then the index is equal to the factor of proportionality. The variable technology cost-of-living index can be decomposed into the product of a constant technology index and a constant price index; indeed, the decomposition can be carried out in two distinct ways, one measuring technology holding prices fixed at P^a and measuring prices holding technology fixed at T^β, the other holding prices at P^b and technology at T^α.[15]

III. The Goods Cost-of-Living Index

The household's preferences for commodities and its technology can be used to define a preference ordering in the goods space. In this section I use this "goods preference ordering" as a basis for welfare evaluation and the construction of a cost-of-living index—the "goods cost-of-living index." Under certain circumstances we

[14] Often we are interested in subindexes for groups of goods or groups of commodities I discussed the construction of such subindexes in the household production framework in an appendix to an earlier version of this paper which is available on request from the Office of Prices and Living Conditions, U.S Bureau of Labor Statistics, Washington, D.C 20212.

[15] There are also a number of properties of the variable technology index which do not depend on restricting prices or technology For example $I(P^a, P^b, Z^o, R, \lambda T^\alpha, T^\beta) = \lambda I(P^a, P^b, Z^o, R, T^\alpha, T^\beta)$ and $I(\lambda P^a, P^b, Z^o, R, T^\alpha, T^\beta) = \lambda I(P^a, P^b, Z^o, R, T^\alpha, T^\beta)$.

may sensibly set out to construct this index, while under others we may construct it inadvertently. In particular, the goods cost-of-living index is precisely the index we obtain if we estimate the goods preference ordering from observed market behavior, and use the estimated preference ordering to construct a cost-of-living index, failing to recognize that production is taking place within the household. Hence, the goods cost-of-living index is of substantial interest.

The goods preference ordering R^* is a derived preference ordering, since households require goods only because goods can produce desired collections of commodities; baskets of goods are evaluated in terms of the best collection of commodities they can produce. Formally, $\hat{X}R^*\bar{X}$ if and only if there exists a $\hat{Z}$ such that $(\hat{Z}, \hat{X}) \in T$ and $\hat{Z}R\bar{Z}$ for all $\bar{Z}$ such that $(\bar{Z}, \bar{X}) \in T$. Hence, the goods preference ordering depends on the household's technology as well as its tastes for commodities, if the reference and comparison technologies are different, they may imply different goods preference orderings even though the household's commodity preferences remain unchanged.[16]

[16]To indicate explicitly the dependence of R^* on R and T, we might write $R^*(R, T)$. The convexity of R^* does not follow from that of R unless we place further restrictions on the technology. In particular, if the production set T is convex, it is easy to show that the convexity of R implies the convexity of R^*. With increasing returns the production set is not convex and it may be so nonconvex as to yield a goods preference ordering which is not convex. For an example, see the author and Wachter, p. 262. The referee correctly points out that there exists a convex goods preference ordering which implies a demand correspondence which contains the demand correspondence generated by the example cited, namely, one whose indifference curves are parallel straight lines. We can define the "goods utility function" $V(X, T)$, by $V(X, T) = \max W(Z)$ subject to $(Z, X) \in T$. That is, $V(X; T)$ is the maximum value of the utility function $W(Z)$ which the household can attain from the goods collection X. The goods utility function is useful in discussing the demand functions for goods implied by the household production model, since maximizing it subject to the budget constraint $\sum p_k x_k = \mu$ yields the goods demand functions. If the goods preference ordering is convex these demand functions exhibit all the properties of conventional demand theory.

The goods cost-of-living index is the index constructed on the basis of the preference ordering R^*. We first define the "goods expenditure function" and then the "goods cost-of-living index."

DEFINITION: The *goods expenditure function* $E^*(P, X^o, R^*, T)$ is defined by

$$E^*(P, X^o, R^*, T) = \min \sum_{k=1}^{n} x_k p_k$$

subject to XR^*X^o

DEFINITION: The *goods cost-of-living index* $\bar{I}^*(P^a, P^b, X^o, R^*, T)$ is defined by

$$\bar{I}^*(P^a, P^b, X^o, R^*, T) = \frac{E^*(P^a, X^o, R^*, T)}{E^*(P^b, X^o, R^*, T)}$$

It follows from the definition of the goods preference ordering that $E(P, Z^o, R, T) = E^*(P, X^o, R^*, T)$ where $(Z^o, X^o) \in T$ and Z^oRZ for all Z such that $(Z, X^o) \in T$; that is, Z^o is the highest indifference curve in the commodity space obtainable from the goods vector X^o.[17] Substituting $E(P, Z^o, R, T)$ for $E^*(P, X^o, R^*, T)$ in the definition of the goods cost-of-living index implies that it is equal to the constant technology cost-of-living index. Formally,

THEOREM 2: *The goods cost-of-living index is equal to the constant technology cost-of-living index.*

$$\bar{I}^*(P^a, P^b, X^o, R^*, T) = \bar{I}(P^a, P^b, Z^o, R, T)$$

where Z^o corresponds to X^o in the sense that $(Z^o, X^o) \in T$ and Z^oRZ for all Z such that $(Z, X^o) \in T$.

We can also write the variable technology cost-of-living index as the ratio of expenditure functions based on goods utility functions, but such a ratio is *not* a cost-of-living index corresponding to a goods preference ordering, since the expenditure func-

[17]This equality also holds when Z^o is replaced by any Z on the indifference curve of Z^o, or when X^o is replaced by any X on the indifference curve of X^o.

tions in the numerator and denominator correspond to different preference orderings in the goods space.

The goods preference ordering is a useful construct if we realize that household production is going on behind the scenes, and are satisfied with the constant technology cost-of-living index. If, however, we fail to recognize that production is taking place within the household and treat the goods preference ordering as the household's true preferences, then we may be misled by the goods cost-of-living index. The difficulty is that the goods preference ordering reflects all changes—whether in the household's technology or in its commodity preference ordering—as changes in the goods preference ordering. Hence, when the comparison technology is superior to the reference technology, using the goods index instead of the variable technology index overstates the change in the cost of living.[18] The result is a natural one: an improvement in technology reduces the variable technology cost-of-living index; but the goods index is equivalent to a constant technology index, which by definition does not reflect the improvement in technology. Thus, with technical progress the goods index misrepresents the welfare position of the household and the compensation required to enable it to maintain a particular standard of living.[19]

IV. Bounds on Exact Indexes

Bounds on cost-of-living indexes are important because we usually do not have enough information to construct exact indexes. In this section and the next I discuss upper bounds on exact indexes in the household production framework.[20] Instead of organizing the discussion of bounding indexes in terms of the exact indexes to which they correspond, I have focused on the information needed to construct them.

In the conventional framework, the Laspeyres index is an upper bound on the cost-of-living index.[21] The Laspeyres index exhibits three important characteristics: (i) it is a fixed weight index defined as the ratio of the costs of base quantities in the reference and comparison situations; (ii) the only information about preferences used to construct it is inferred from the fact that the base quantities were chosen in the reference situation; and (iii) it is an upper bound on the exact cost-of-living index. When we go from the conventional framework to the household production model there are two Laspeyres-type indexes which exhibit these characteristics. One of these, the "goods Laspeyres index," is the ratio of the costs of a particular collection of market goods: the other, the "commodity Laspeyres index," is the ratio of the costs of a particular collection of commodities.

Since inadequate information provides the motivation for constructing bounds on exact indexes, it is useful to recognize a spectrum of cases of information availability. At one pole we have full or complete information about the household's tastes and technology, so that we can construct exact indexes.[22] At the other pole,

[18]Suppose the household's technology improves so that the expenditure required to attain any particular indifference curve of the commodity preference ordering using the comparison technology is less than the expenditure required using the reference technology. Then $E(P, Z^o, R, T^a) > E(P, Z^o, R, T^b)$. Hence,

$$I(P^a, P^b, Z^o, R, T^a, T^b) = \frac{E(P^a, Z^o, R, T^a)}{E(P^b, Z^o, R, T^b)} < \frac{E(P^a, Z^o, R, T^b)}{E(P^b, Z^o, R, T^b)} = \bar{I}^*(P^a, P^b, X^o, R^*, T^b)$$

where Z^o corresponds to X^o in the sense that $Z^o R Z$ for all Z such that $(Z, X^o) \in \mathscr{T}$. Similarly, $I(P^a, P^b, Z^o, R, T^a, T^b) < \bar{I}^*(P^a, P^b, X^o, R^*, T^a)$ where Z^o corresponds to X^o.

[19]Fisher and Shell, p. 104, reach essentially the same conclusion in their discussion of "quality change" (read, "technical progress") and "taste change" (focusing on the goods cost-of-living index) without formally introducing the machinery of the household production model.

[20]The analysis of lower bounds is left to the reader.

[21]A precise statement is given in Section I.

[22]If we are only interested in constructing the exact index corresponding to a given Z^o, then we need to know only the indifference curve of Z^o, not the entire indifference map. Similarly, if we are only interested in the index comparing P^a and P^b, we need not know

"extreme ignorance," we know only the prices in the reference and comparison situations. In this case, we can conclude that the exact index is bounded above by the largest percentage price increase; furthermore, this is the "best bound" which can be established given the limited information available.[23] Between these poles are a spectrum of intermediate cases. The closest analogue of the traditional Laspeyres case is one in which we know nothing about the household's tastes or technology beyond what we can infer from the goods collection it purchased in the reference situation; that is, we have no information about household production which may or may not be going on behind the scenes. Using only information about the goods collection purchased in the reference situation, I define the goods Laspeyres index.

DEFINITION: The *goods Laspeyres index* $\bar{J}(P^a, P^b, X^b)$ is defined by

$$\bar{J}(P^a, P^b, X^b) = \sum_{k=1}^{n} x_k^b p_k^a \Big/ \sum_{k=1}^{n} x_k^b p_k^b$$

where X^b is the collection of goods the household purchased when facing prices P^b with technology T^b and income μ^b.

A version of the usual Laspeyres bounding theorem holds and follows directly from the type of argument used to establish the Laspeyres bounding theorem without household production.

THEOREM 3: *The goods Laspeyres index is an upper bound on the constant technology cost-of-living index:*

$$\bar{I}(P^a, P^b, X^b, R^b, T^b) \leqq \bar{J}(P^a, P^b, X^b)$$

where X^b is the collection of goods which permits the household to attain the indifference curve of X^b with minimum expenditure when facing goods prices P^b with technology T^b.[24]

The requirement that the index is based on the technology of T^b, the technology corresponding to P^b, is crucial to this result. This condition augments but does not replace the traditional Laspeyres stipulation that the cost-of-living index is based on R^b and evaluated at X^b.[25]

Another intermediate position between the polar cases of complete information and extreme ignorance is one in which technology is known but tastes are not.[26] This is an especially attractive assumption because the usefulness of the household production model is substantially enhanced when we can observe both goods and commodities and infer the household's technology from these observations. When our observations are limited to the household's purchases of market goods and we never observe the commodities it produces with them, the household production model is indistinguishable from the conventional theory of household behavior, even if we have enough data on household purchases of market goods to enable us to infer the goods demand functions and the goods preference ordering.[27]

If we know the household's technology, and if we know the commodity collection the household chose in the reference situa-

the entire technology or cost function, but only the values of the cost function corresponding to P^a and P^b

[23]See the author (1971, sec 4) for a discussion of best bounds and the preference orderings to which they correspond in the traditional framework.

[24]This bounding theorem does not depend on the convexity of the goods preference ordering.

[25]The goods Laspeyres index does not permit us to draw any conclusions about the variable technology cost-of-living index, but if we also know that the comparison technology "dominates" the reference technology—in the sense that any collection of commodities can be produced with the comparison technology using no more inputs than were required by the reference technology—then the goods cost-of-living index is also an upper bound on the variable technology cost-of-living index.

[26]We could also consider the intermediate case in which tastes are known and technology is not, but such a case is less likely and therefore less interesting

[27]See the author and Wachter for a discussion of this point.

tion, then we can construct the commodity Laspeyres index

DEFINITION: The *commodity Laspeyres index* $J(P^a, P^b, Z^b, T^a, T^b)$ is defined by

$$J(P^a, P^b, Z^b, T^a, T^b) = \frac{C(P^a, Z^b; T^a)}{C(P^b, Z^b; T^b)}$$

This is a Laspeyres-type index in the sense that it is based on a fixed commodity consumption pattern Z^b; it can be calculated without knowledge of the household's tastes (beyond what can be inferred from the fact that Z^b was chosen in the reference situation), but it does require knowledge of the technology. The index reflects changes in the minimum cost of producing the fixed commodity bundle Z^b, and knowledge of the household's technology is needed to calculate the cost-minimizing input combinations in different price-technology situations. Like the Laspeyres index, it does not allow for changes in the commodity consumption pattern induced by such changes The usual Laspeyres argument shows that it is an upper bound on the variable technology cost-of-living index

THEOREM 4: *The commodity Laspeyres index is an upper bound on the variable technology cost-of-living index*

$$I(P^a, P^b, Z^b, R^b, T^a, T^b) \leq J(P^a, P^b, Z^b, T^a, T^b)$$

where Z^b is the collection of commodities consumed in the price-technology-expenditure situation (P^b, T^b, μ^b).

V. The Role of Implicit Commodity Prices

"Implicit commodity prices" play a major role in the household production literature. In this section I examine their usefulness in welfare evaluation and the construction of cost-of-living indexes. I argue that they have no real role to play in the construction of exact indexes, although there are some important special cases (for example, constant returns together with the absence of joint production) in which exact indexes can be written in terms of implicit commodity prices. But implicit commodity prices can play a major role in the construction of bounds. In particular, if the household's technology is known, and if the income-consumption curve in the commodity space corresponding to the comparison price-technology situation is known, then commodity shadow prices can be used to establish a better bound on the variable technology index than that provided by the commodity Laspeyres index.

The implicit or shadow price of a commodity is its marginal cost. Formally,

DEFINITION: The *implicit price* of commodity r, $\pi_r(P, Z, T)$, is defined by[28]

$$\pi_r(P, Z; T) = \frac{\partial C(P, Z; T)}{\partial z_r}$$

In general, implicit commodity prices depend not only on the household's technology and goods prices, but also on the *quantities of all commodities consumed by the household* (see the author and Wachter and W. E. Diewert, sec. 7). Implicit commodity prices are independent of the particular collection of commodities consumed by the household (and thus independent of the household's tastes) if and only if the technology exhibits constant returns to scale and no joint production.[29] In this case we can write the cost function as

$$C(P, Z; T) = \sum_{i=1}^{m} C^r(P, 1; T) z_i$$

where $C^r(P, z_r; T)$ is the cost function associated with the production of commodity r, implicit commodity prices are equal to the unit cost functions, $C^r(P, 1; T)$.

The commodity price cost-of-living index is defined by imagining that the household

[28] If the cost function is not differentiable at a particular point, implicit commodity prices are undefined there

[29] A proof can be found in the author and Wachter where it is argued that this is the only case in which implicit commodity prices are useful explanatory variables in demand analysis.

purchases commodities instead of producing them, and that it is a price taker in the market for commodities. To emphasize the distinction between the implicit commodity prices $\pi(P, Z; T)$, defined as the partial derivatives of the cost function, and the predetermined commodity prices used to define the commodity price cost-of-living index, we denote the latter by $\bar{\pi}$. We begin by defining the commodity price expenditure function.

DEFINITION: The *commodity price expenditure function* $E^{**}(\bar{\pi}, Z^o, R)$ is defined by

$$E^{**}(\bar{\pi}, Z^o, R) = \min \sum_{i=1}^{m} z_i \bar{\pi}_i$$

subject to ZRZ^o

DEFINITION: The *commodity price cost-of-living index* $I^{**}(\bar{\pi}^A, \bar{\pi}^B, Z^o, R)$ is defined by

$$I^{**}(\bar{\pi}^A, \bar{\pi}^B, Z^o, R) = \frac{E^{**}(\bar{\pi}^A, Z^o, R)}{E^{**}(\bar{\pi}^B, Z^o, R)}$$

The commodity price cost-of-living index is defined without reference to goods prices or the household's technology; it begins with commodity prices, which it takes as given. The resulting index satisfies all of the theorems of traditional cost-of-living index theory, where the prices referred to in the theorems are understood to be the predetermined commodity prices, $\bar{\pi}$.

If the household's technology exhibits constant returns to scale and no joint production, then implicit commodity prices are uniquely determined by goods prices and the household's technology, and the cost function can be written in terms of implicit commodity prices as $C(P, Z; T) = \Sigma \pi_i(P; T) z_i$.

It is easy to verify that in this case the variable technology cost-of-living index coincides with the commodity price cost-of-living index when commodity prices are evaluated at corresponding price-technology situations. That is, under constant returns to scale and no joint production[30]

$$I^{**}(\bar{\pi}^A, \bar{\pi}^B, Z^o, R) = I(P^a, P^b, Z^o, R, T^\alpha, T^\beta)$$

where $\bar{\pi}^A = \pi(P^a; T^\alpha)$
and $\bar{\pi}^B = \pi(P^b; T^\beta)$

We now turn to the role of commodity prices in calculating bounding indexes. I first show that it is sometimes possible to write the commodity Laspeyres index in terms of implicit commodity prices, even when the variable technology index cannot be rewritten in this way. This result is of limited interest, however, because it is difficult to imagine circumstances under which we would know commodity shadow prices

[30]The case of constant returns to scale and no joint production is not the only one in which the commodity price cost-of-living index coincides with the variable technology index Muellbauer shows that even with joint production, they coincide when the reference and comparison technologies are homogeneous of the same degree But in this case, the implicit commodity price formulation is no more transparent than the cost function-expenditure function formulation of Section II Note that to calculate the exact index, implicit commodity prices in the numerator (denominator) must be evaluated at $Z^{oa\alpha}(Z^{ob\beta})$, the least cost commodity basket which attains the indifference curve of Z^o at prices $P^a(P^b)$ with technology $T^\alpha(T^\beta)$ Muellbauer suggests that the commodity shadow price formulation requires less information than the cost-expenditure approach, but I do not find his position persuasive He assumes that implicit commodity prices of the reference and comparison situations are known or observable, while the household's technology is not But if we knew the implicit commodity price *functions*, we could find the associated cost functions and from them the entire technology Muellbauer assumes that we do not know the implicit commodity price functions, but only the value of these functions at the points $(P^a, Z^a; T^a)$ and $(P^b, Z^b; T^b)$. However, he does little to motivate his assumptions about information Since implicit commodity prices are not directly observable, they must be calculated using estimates of the parameters of the household's technology. Once a functional form for the technology is specified, its parameters can be estimated from observations on inputs and outputs or on outputs and cost The parameter estimates can then be used to calculate not only commodity shadow prices at Muellbauer's points, but the implicit commodity prices corresponding to any configuration of goods prices and commodity outputs. Estimating the parameters of the household's technology is a prerequisite to calculating commodity shadow prices

but not have enough information to construct the commodity Laspeyres index. I then show that if we have some additional information about tastes it may be possible to use implicit commodity prices to construct a better bound on the variable technology index than that provided by the commodity Laspeyres index.

I begin by defining a family of commodity Laspeyres indexes.

DEFINITION: The *commodity price Laspeyres index* $J^{**}(\pi^A, \pi^B, Z^b)$ is defined by

$$J^{**}(\pi^A, \pi^B, Z^b) = \frac{\sum_{i=1}^{m} z_i^b \pi_i^A}{\sum_{i=1}^{m} z_i^b \pi_i^B}$$

I describe this as a family of indexes rather than a single index because its interpretation and usefulness depend on the further specification of the commodity prices; in particular, I do not require $\pi^A(\pi^B)$ to be the commodity prices corresponding to the comparison (reference) price-technology-consumption situation. When Z^b is the collection of commodities which attains the indifference curve of Z^b at minimum cost when evaluated at the commodity prices π^B, then the commodity price Laspeyres index is an upper bound on the commodity price cost-of-living index $I^{**}(\pi^A, \pi^B, Z^b, R) \leq J^{**}(\pi^A, \pi^B, Z^b)$. This follows from the usual Laspeyres argument, but the result is of limited interest because the two indexes treat commodity prices as exogenous.

In certain cases it is possible to rewrite the commodity Laspeyres index in terms of implicit commodity prices. In particular, suppose we have no information about preferences beyond what can be inferred from the commodity bundle chosen in the reference situation, but that the household's reference and comparison technologies are known and the implied cost functions satisfy the differential equation $C(P, Z; T) = \gamma(Z)\Sigma z_i \partial C_i(P, Z; T)/\partial z_i$ where $\gamma(Z)$ is the same for both technologies. Then the commodity price Laspeyres index is equal to the commodity Laspeyres index and, hence, is an upper bound on the variable technology cost-of-living index

$$I(P^a, P^b, Z^b, R^b, T^a, T^b) \leqq J(P^a, P^b, Z^b, T^a, T^b) = J^{**}(\pi^A, \pi^B, Z^b)$$

where implicit commodity prices are evaluated at $\pi^A = \pi(P^a, Z^b; T^a)$ and $\pi^B = \pi(P^b, Z^b; T^b)$. This result depends on the fact that the $\gamma(Z)$'s cancel when the cost functions in the numerator and denominator are evaluated at the same value of Z. For example, consider any pair of technologies which can be viewed as two-stage processes in which at the first stage market goods are transformed into a single homogeneous output q, according to a constant returns to scale technology, and, at the second stage the homogeneous output is used to produce commodities. Processes of this type imply cost functions of the form $C(P, Z; T) = \psi(P; T)\phi(Z; T)$ where $\psi(P; T)$ is the unit cost function corresponding to the first-stage technology and $q = \phi(Z; T)$ implicitly defines the second stage. If the second-stage processes are identical for the reference and comparison technologies, then $\phi(Z; T^a) = \phi(Z; T^b) = \phi(Z)$ and the cost functions are of the required form.[31]

Neither of the bounding indexes just described yields a better bound on the exact indexes than could be obtained without implicit commodity prices, they are merely restatements in terms of commodity shadow prices of results obtained in Section IV. I now show that it is sometimes possible to use implicit commodity prices in conjunction with some additional information to obtain better results. In particular, we can improve on the commodity Laspeyres index if, in addition to knowing the

[31]This result is a substantial generalization of a theorem of Muellbauer Muellbauer's "consistency condition," p. 983, which he contends is necessary if the commodity price Laspeyres index is to make sense, is unduly strong, and it is not satisfied by the two-stage technology described above. Unfortunately, this result is only interesting when we know implicit commodity prices but not the household's technology.

household's technology, we also know the income-consumption curve in the commodity space corresponding to the comparison price-technology situation (P^a, T^a). It is assumed that preferences are convex and are the same in both situations.

The bounding argument combines two observations: first, if a commodity collection Z^* lies on a higher indifference curve than Z^b, then giving the household enough income to produce Z^* will at least compensate it for any price or technological changes which have taken place. Second, suppose that preferences are convex and that Z^* is chosen from a feasible set; if Z^b lies on the "pseudo budget set" defined by the hyperplane tangent to the feasible set at Z^*, then Z^* is "implicitly revealed preferred" to Z^b. This is substantially stronger than the obvious revealed preference assertion that if Z^* is chosen from a feasible set, then it is at least as good as any bundle in the feasible set. A stronger conclusion (stronger, at least, when the feasible set is convex) is possible when preferences are convex and the boundary of the feasible set is not a hyperplane With convex preferences, a "pseudo budget line" tangent to the feasible set at the chosen bundle is also tangent to the indifference curve at this bundle, so the chosen bundle is at least as good as every bundle in the pseudo budget set.[32] In particular, if Z^b lies in the pseudo budget set corresponding to Z^*, then the variable technology cost-of-living index must be less than $C(P^a, Z^*; T^a)/C(P^b, Z^b; T^b)$. Whether or not this is a better bound than that implied by the commodity Laspeyres index depends on the location of Z^* relative to Z^b. However, if the technology is strictly convex and Z^b happens to lie *on* the pseudo budget line corresponding to Z^*, then this will be a better bound than the commodity

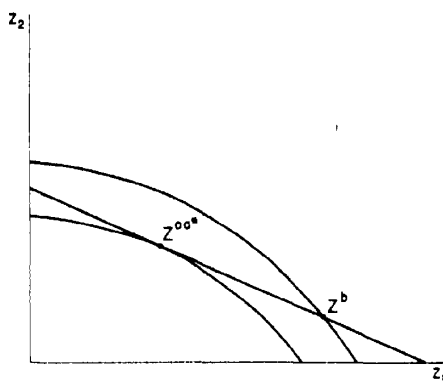



FIGURE 1

Laspeyres index.[33,34]

If we know the income-consumption path in the commodity space corresponding to (P^a, T^a), then we can use the above argument systematically to bound the variable technology cost-of-living index. The procedure works by finding a commodity bundle Z^{oa*} on the income-consumption path of (P^a, T^a) such that Z^b lies on the boundary of the pseudo budget set corresponding to that commodity bundle. The commodity Laspeyres index reflects the cost of the feasible set whose boundary passes through Z^b (see Figure 1).

Formally, let Z^{a*} be a point on the income-consumption path corresponding to (P^a, T^a, μ^{a*}). Then there is a shadow price system $\{\pi_i(P^a, Z^{a*}; T^a)\}$ associated with this point, and we define "implicit expenditure," $\tilde{\mu}(P^a, Z^{a*}; T^a)$ by $\tilde{\mu}(P^a, Z^{a*}; T^a) = \Sigma z_i^{a*} \pi_i(P^a, Z^{a*}; T^a)$. By the previous argument, Z^{a*} is at least as good as all Z in the pseudo budget set defined by $\Sigma z_i \pi_i(P^a, Z^{a*}; T^a) \leq \tilde{\mu}(P^a, Z^{a*}; T^a)$. Let $S(Z^b)$ denote the set of all Z^{a*} such that $\Sigma z_i^b \pi_i(P^a, Z^{a*}; T^a) \leq \tilde{\mu}(P^a, Z^{a*}; T^a)$. Then $Z^{a*} R Z^b$ for all $Z^{a*} \in S(Z^b)$ and

[32] If the boundary of the feasible set is itself a hyperplane (as it is with constant returns and no joint production), then the pseudo budget set coincides with the feasible set. But if the feasible set is strictly convex, then the pseudo budget set permits us to identify some commodity bundles which were not in the feasible set as being no better than Z^a.

[33] Provided Z^* and Z^b are distinct

[34] In the traditional framework without household production the boundary of the feasible set is a hyperplane, and knowing the collection of goods purchased in the comparison price-income situation, or even the entire income-consumption curve, does not enable us to improve on the Laspeyres bound.

$C(P^a, Z^{oa*}; T^a) \leq C(P^a, Z^{a*}; T^a)$ for all $Z^{a*} \in S(Z^b)$. Hence,[35,36]

$$I(P^a, P^b, Z^b, R^b, T^a, T^b) \leq \frac{C(P^a, Z^{oa*}; T^a)}{C(P^b, Z^b, T^b)}$$

VI. Conclusion

The evaluation of welfare and the construction of a cost-of-living index in the household production framework require a careful specification of the alternative situations being compared. The variable technology cost-of-living index measures the total effect on the household of changes in goods prices and changes in its technology; the constant technology index is independent of changes in the household's technology and reflects only changes in goods prices. The first three sections of this paper discussed exact cost-of-living indexes under the assumption that the household's tastes and technology were known. Bounds on exact indexes are important because this complete information is usually not available. When neither the household's tastes nor its technology is known, we can calculate the goods Laspeyres index which is an upper bound on the constant technology index. When the technology is known but tastes are not, we can calculate the commodity Laspeyres index which is an upper bound on the variable technology index. When both the technology and the expansion path corresponding to the comparison price-technology regime are known, we can calculate commodity shadow prices and use them to construct a Laspeyres-type index which provides a better bound on the variable technology index than the commodity Laspeyres index.

Striking similarities exist between the cost-of-living index in the household production framework and two other generalizations of the cost-of-living index from its traditional setting. Both subindexes of the cost-of-living index (see the author, 1975a, and Charles Blackorby and Robert Russell) and the "social cost-of-living index" based on a Bergson-Samuelson social welfare function (see the author, 1976), like household production indexes, involve two-stage maximization.[37] In the household production framework, one stage corresponds to the household's technology and the other to its tastes; with subindexes, the two stages reflect the household's preferences for goods within each group and for aggregates of goods, and with social cost-of-living indexes, one represents the household's preferences and the other the social preferences expressed by the Bergson-Samuelson social welfare function. In all three, alternative assumptions about the in-

[35] Muellbauer, pp. 985–86, argues that if the household's preferences are homothetic and identical in the reference and comparison situations, and the household's technology in both periods exhibits constant returns to scale, then the variable technology cost-of-living index is bounded by the commodity price Laspeyres index where $\pi^A = \pi(P^a, Z^a, T^a)$ and $\pi^B = \pi(P^b, Z^b, T^b)$. This result follows from the general procedure just described. Under his assumptions, the income-consumption path corresponding to (P^a, T^a) is a ray from the origin, and hence a single commodity bundle on the path identifies the entire path. With constant returns to scale, implicit commodity prices are constant along such a ray and implicit expenditure at a point on the income-consumption path is equal to the cost of that commodity bundle.

[36] If we know the commodity bundle purchased in the comparison situation, (P^a, T^a, μ^a), but not the entire income-consumption curve, we may still be able to establish a better bound on the variable technology cost-of-living index than that provided by the commodity Laspeyres. A prerequisite, however, is that Z^b lie in the pseudo budget set corresponding to Z^a. If this condition is satisfied, we can infer that $C(P^a, Z^a; T^a)/C(P^b, Z^b, T^b) = \mu^a/\mu^b$ is an upper bound on the variable technology index. Of course $C(P^a, Z^a, T^a)$ need not be smaller than $C(P^a, Z^b, T^a)$, so this need not be a better bound than the commodity Laspeyres index. Our general conclusion is that provided Z^b lies in the pseudo budget set of Z^a—the variable technology index is bounded above by the minimum of μ^a/μ^b and the commodity Laspeyres index. If we also know additional points on the income-consumption curve corresponding to (P^a, T^a), we can use them in the same way to attempt to improve on the commodity Laspeyres index.

[37] In Blackorby and Russell separability of the expenditure function is a prerequisite for defining subindexes. In the author (1975a), subindexes are defined without assuming separability, but they are better behaved when the direct utility function is separable.

formation available for index construction permit the development of a variety of indexes which bound the exact index.

REFERENCES

G. S. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

C. Blackorby and R. R. Russell, "Indices and Subindices of the Cost of Living and the Standard of Living," *Int. Econ. Rev.*, forthcoming.

W. E. Diewert, "Walras' Theory of Capital Formation and the Existence of a Temporary Equilibrium," in Gerhard Schwödiauer, ed., *Equilibrium and Disequilibrium in Economics*, Dordrecht 1976.

F. M. Fisher and K. Shell, "Taste and Quality Change in the Pure Theory of the True Cost-of-Living Index," in J. N. Wolfe, ed., *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, Edinburgh 1968, 97–139.

Kelvin J. Lancaster, (1966a) "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.

———, (1966b) "Change and Innovation in the Technology of Consumption," *Amer. Econ Rev Proc.*, May 1966, *56*, 14–23.

———, *Consumer Demand, A New Approach*, New York 1971.

R. T. Michael and G. S. Becker, "On the New Theory of Consumer Behavior," *Swedish J. Econ.*, Dec. 1973, *75*, 378–96.

J. Muellbauer, "Household Production Theory, Quality and the 'Hedonic Technique'," *Amer. Econ. Rev.*, Dec. 1974, *64*, 977–94.

Louis Phlips, *Applied Consumption Analysis*, Amsterdam 1974.

——— **and R. Sanz-Ferrer**, "A Taste-Dependent True Index of the Cost of Living," *Rev. Econ. Statist.*, Nov. 1975, *57*, 495–501.

R. A. Pollak, "The Theory of the Cost of Living Index," res. disc. paper no. 11, Res. Div., Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, June 1971

———, (1975a) "Subindexes of the Cost of Living Index," *Int. Econ. Rev.*, Feb. 1975, *16*, 135–50.

———, (1975b) "The Treatment of 'Quality' in Demand Analysis and the Cost of Living Index," mimeo., Aug. 1975.

———, "The Social Cost of Living Index," disc. paper no. 378, Univ. Pennsylvania, June 1976.

——— **and M. L. Wachter**, "The Relevance of the Household Production Function and Its Implications for the Allocation of Time," *J. Polit Econ.*, Apr. 1975, *83*, 255–77.

Uncertain Externalities, Liability Rules, and Resource Allocation

By PETER H. GREENWOOD AND CHARLES A. INGENE*

The rich literature on the economics of externalities has been confined to the analysis of what we term certainty—perfect knowledge of the impact an action taken by one economic unit will have on another unit. The burden of these studies has been that "... if market transactions were costless, all that matters (questions of equity apart) is that the rights of the various parties should be well-defined and the results of of legal actions easy to forecast" (Ronald Coase, p. 19), since "... the affected parties might engage in bargaining and attempt to arrange a solution between themselves" (Otto Davis and Andrew Whinston, p. 113). So long as negotiations (market transactions) are costless, the allocation of resources at the conclusion of the bargaining process is socially optimal because it has the same characteristics as the equilibrium position attained by a merger of the affected parties into a single firm which fully internalizes externalities. This powerful conclusion, independence from the assignment of liability by the legal system (questions of equity apart), has come to be called the "Coase Theorem."

We extend the results of our precursors by analyzing an uncertain distribution of externalities, postulating a situation in which one firm's activities affect another firm in a random fashion. The existence of this uncertainty is sufficient to cause any firm facing it to modify its behavior in a subtle but significant manner: maximization is of the expected utility from profit rather than of profit itself. This response to an inevitably risky situation enables the firm to incorporate into its maximization problem both the distribution of the externality and its own attitude toward risk.

We demonstrate that given uncertainty, the allocation of resources may not be independent of legal liability; thus, the outcomes of bargaining and of merger may not coincide. Outcomes independent of liability and the equivalence of bargaining and merger are dependent both upon the existence of a stock market in which risk may be shared and upon the absence of indivisibilities in wealth. Although the traditional conclusions under certainty are correct and are contained within our model as a special case, in an uncertain world the Coase Theorem is valid only with the more stringent assumptions stated above.

In Section I we establish our mathematical model, a Taylor series expansion of the expected utility from profit, and use it to discuss the result of a merger by the involved parties. The series leads to the Pratt-Arrow measure of absolute risk aversion. We show that the equilibrium position "fully" internalizes the externality; but, the definition of fully may be dependent upon the risk attitude of whoever controls the merged firm. We also discuss the fact that a dominant shareholder or an autonomous manager may prevent the complete spreading of risk through a stock market.

The second section deals with legally permissible, uncompensated externalities. The affected firm will then approach the polluter in a costless "private bribery market" to seek an alteration in the level of the externality creating output.[1] The polluter

*Assistant professor of resource economics, University of New Hampshire; and assistant professor of economics, University of Oklahoma, respectively. We are grateful to G. Borts, R. Dacey, and an anonymous referee for their helpful comments. None of them are sufficiently risk preferring to wish to share responsibility for an article which will have an uncertain reception; thus, common law allocates liability for remaining errors to the authors.

[1] For ease of exposition, we will discuss only a negative externality, our mathematics is perfectly general, however. We do assume that pollution and output are directly related in order to avoid the complication of

will comply so long as the usual, profit-maximizing, marginal conditions are met. At the completion of the bargaining process there will be no gains from trade to further bargaining; the achieved optimum will reflect the risk attitude of the firm which is effectively responsible for damages. We also show that a risk preferrer and a risk averter demand different levels of output reduction in the bribery market. In Section III the polluter is held legally responsible; the results parallel those in Section II, with the polluter's risk attitude incorporated into the equilibrium conditions.

In Section IV we show that the government, by use of lump sum or per unit taxes (subsidies), can alter the allocation of resources; therefore, it can correct any bribery market failure which it perceives. A per unit tax (subsidy) is shown to possess an "income effect" and a "substitution effect" upon output levels whereas a lump sum tax (subsidy) has an income effect only. We state our conclusions in Section V. In the Appendix, Section A, we investigate the sensitivity of the equilibrium output levels to shifts in the probability distribution of damages. In particular, we examine an additive shift of the mean, a multiplicative shift of the variance, and a proportional shift of all moments. Each has an income effect on output levels that is akin to the effect on output of a change in fixed costs. In addition, the last case has a substitution effect. The sign of the change in output levels is shown to be related to whether risk aversion is increasing, invariant, or decreasing with wealth. In the Appendix, Section B, we develop a market which permits the sharing of risk between the polluter and the pollutee, and show that in the absence of a market for perfect risk sharing, the Coase Theorem does not hold.

I. Pareto Optimality

Consider a pair of firms, A and B, which are ineluctably linked via an externality generated by A, randomly impacting upon B. A neoclassical example of a (negative) externality is a factory which pollutes the air; this "side-product" harms a nearby washerwoman; the degree of damage is dependent upon imperfectly predictable climatic conditions.

Firm A's profits are a function of its own output and a set of known parameters, simply written as $\Pi(Q)$. We assume Π to be a continuous, twice-differentiable function with a unique maximum. Firm B's profits are a function of the outputs of both firms, written as $\pi(Q,q)$. Uncertainty is introduced by allowing the impact of Q upon π to vary stochastically. The π is distributed as $(\bar{\pi}, v)$, where $\bar{\pi}$ is the expected value of π and v is the variance of profits about the mean.[2] For expositional ease we will treat the externality as negative; the mathematics imposes no such constraint. Both firms are regarded as price takers in all markets; they need not be perfect competitors.

A generally accepted economic doctrine is that full internalization of the externality is necessary if Pareto optimality is to obtain. The (conceptually) simplest method of accomplishing this is a merger of A and B. However, even this merged firm cannot know a priori the exact effect its output decisions will have on profits; it must consider myriad possibilities, weighting them on the basis of their likelihood and desirability. Thus, the merged firm will maximize its expected utility of profits.

We utilize a Taylor series to describe the utility function.[3,4] While the series may be

inferior factors which Charles Plott has noted. Our model can be extended to incorporate such factors.

[2]We assume that there exists a known relationship between Q and a physical measure of "pollution at the smokestack" which is independent of the state of nature. Uncertainty occurs as "pollution enters the air." Both $\bar{\pi}$ and v are functions of Q and q.

[3]Otto Loistl has shown that this assumption is not as innocent as it appears at first blush. However, Kenneth Arrow has argued that for a utility function to satisfy the von Neumann-Morgenstern axioms it must be bounded from above. A Taylor series expansion of such a utility function will converge to the function; therefore, our use of a Taylor series is legitimate. Note that truncating the series at, say, the fourth-order term would complicate our mathematics without materially affecting the results.

[4]To avoid the possibility of cyclical majorities, decisions are made by an individual. At this time the

expanded about any arbitrary values, wise choices ought to lead to an economically enlightening conclusion. We believe such numbers to be Π^* and $\bar{\pi}^*$. They are the optimal level of profits from part A and the optimal level of expected profits from part B of the merged firm. We start by expanding the utility of profits:

$$(1)\quad u(\Pi + \pi) = u(\cdot) + u'(\cdot)\{(\Pi - \Pi^*) + (\pi - \bar{\pi}^*)\} + (1/2)u''(\cdot)\{(\Pi - \Pi^*)^2 + 2(\Pi - \Pi^*)(\pi - \bar{\pi}^*) + (\pi - \pi^*)^2\} + 0^{(3)}$$

Primes denote derivatives, $(\cdot) \equiv (\Pi^* + \bar{\pi}^*)$, and $0^{(3)}$ denotes terms of order 3 and above. We truncate the series by treating $0^{(3)}$ as negligible. Retention of the remainder would complicate the analysis without affecting its essential components. Take the expected value of (1), noting that the variance is defined as $E[(\pi - \bar{\pi})^2] = E[\pi^2] - \bar{\pi}^2$ and that the expansion of $u(\Pi + \bar{\pi})$ about Π^* and $\bar{\pi}^*$ (rather than $u(\Pi + \pi)$ about Π^* and $\bar{\pi}^*$, as in equation (1)) is very similar to (1). It follows that the expected value of (1) may be expressed as

$$(2)\quad E[u(\Pi + \pi)] = u(\Pi + \bar{\pi}) + (1/2)u''(\cdot)v$$

In (2), $(1/2)u''(\cdot)$ is a specific number while v and $u(\Pi + \bar{\pi})$ are functions of both Q and q.

The merged firm will maximize (2) by setting the partial derivatives to zero (throughout this paper we assume that second-order conditions are satisfied)

$$(3a)\quad u'(\Pi + \bar{\pi})(\Pi_Q + \bar{\pi}_Q) + (1/2)u''(\cdot)v_Q = 0$$

$$(3b)\quad u'(\Pi + \bar{\pi})(\bar{\pi}_q) + (1/2)u''(\cdot)v_q = 0$$

a subscript denotes the partial derivative with respect to the argument (i.e., $\partial\Pi/\partial Q \equiv \Pi_Q$). When equations (3) hold, $\Pi = \Pi^*$ and $\bar{\pi} = \bar{\pi}^*$, so we may rewrite (3) as:

$$(4a)\quad \Pi_Q + \bar{\pi}_Q = -\frac{1}{2}\frac{u''(\cdot)}{u'(\cdot)}v_Q = \frac{1}{2}r(\cdot)v_Q$$

$$(4b)\quad \bar{\pi}_q = -\frac{1}{2}\frac{u''(\cdot)}{u'(\cdot)}v_q = \frac{1}{2}r(\cdot)v_q$$

where $r(\cdot)$ is the Pratt-Arrow measure[5] of absolute risk aversion evaluated at profit level $(\cdot)$. Since profit is a "good," $u'(\cdot) > 0$. The sign of $u''(\cdot)$ defines the attitude towards risk. If it is positive the firm is a risk preferrer; if zero, the firm is risk neutral. Risk aversion is defined as $u''(\cdot) < 0$. Obviously, $r(\cdot)$ is of the opposite sign from $u''(\cdot)$.

The firm chooses the output combination $\{Q^*, q^*\}$ which satisfies equations (4). Thus, it sets its marginal profit from each type of output (net of damages) equal to one-half $r(\cdot)$ times the rate of change of the variance of π with respect to output.[6] Call $(1/2)r(\cdot)v_i$, $i = Q, q$, the firm's adjusted risk attitude. If in equilibrium $v_i = 0$, the firm's inclination towards risk is irrelevant; it acts as if it were an expected profit maximizer. This can occur if v has a maximal value, if the randomness of π is unrelated to output (known as system uncertainty), or if the variance itself is zero. Thus, the traditional analysis of certainty is contained within our model as a special case.

Because equations (4) reflect the utility function of the decision maker, we ask if output levels are dependent upon whoever controls the corporation. In particular, if the manager of part A were elevated to control, his measure of absolute risk aversion $R(\cdot)$ would replace $r(\cdot)$ in equations (4). Does $R(\cdot) = r(\cdot)$? A further question concerns nationalization, or at least governmental intervention: what is $S(\cdot)$, society's attitude towards risk? If $S(\cdot) \neq R(\cdot)$, $r(\cdot)$, society may wish to intervene in the allocative process.

Arrow and Robert Lind argue that if the profits of the (merged) firm are statistically independent of other components of national income, if there is no corporate in-

decision maker is the manager of part B, the expansion is of his utility function.

[5]See Arrow or John Pratt.

[6]If Q or q is zero, $v = 0$. Thus both v_Q and v_q must be positive over some output range, this is Hayne Leland's "principle of increasing uncertainty." However, saturation levels of pollution may be reached, thus v_Q and v_q may become nonpositive at sufficiently high output levels

come tax, if there are a large number of shareholders (each holding a small portion of his wealth in this firm), and "... if managers were acting in the interest of the firm's shareholders, they would essentially ignore risks ..." (p. 376). Thus, $r(\cdot) = 0 = R(\cdot)$. They also argue that the government should act in the same manner ($S(\cdot) = 0$). Clearly, if the Arrow-Lind assumptions hold we have certainty equivalence and the Coase Theorem always holds.

However, indivisibilities may negate these results. An autonomous manager who receives a significant segment of his income from the firm will obey his own, not the market's, risk measure. Similarly, if "... in order to control the firm, some shareholder [holds] a large block of stock which is a significant component of his wealth" (Arrow-Lind, p. 376), the firm should use his, not the market's, risk measure. With indivisibilities, corporate control matters. The externality will always be fully internalized, although the sense of fully will not always be the same. Society may then wish to intervene to obtain a Pareto optimal resource allocation.

Prior to discussing potential governmental intervention, we investigate liability rules under which the independent firms A and B may bargain privately to lessen the impact of the externality. Before bargaining can occur, "[I]t is necessary to know whether the damaging business is liable or not for damage caused since without the establishment of this initial delimitation of rights there can be no market transactions to transfer and recombine them" (Coase, p 8).

II. Legally Permissible Pollution

Consider the case of a legal system freely granting the right of unlimited generation of an externality to firm A. Firm B is injured by the externality and, therefore, has cause to seek an improvement in its own situation.[7] While protective measures such as physical movement of its plant or alteration of its productive process to a less affectable technology are possibilities, they will not concern us here. We are interested in the extent to which the two firms may bargain to their mutual advantage—firm B by obtaining a profitable reduction in Q and firm A by profitably reducing its own output.

For every unit reduction in Q, firm B's profits rise; it follows that B must be willing to pay some sum of money, not exceeding its marginal profit gain, to obtain a unit reduction in Q. In fact, B must have a demand curve for a reduction in Q, a curve which measures the marginal benefit to B of such a reduction. If the loss to B rises at an increasing rate with added units of A's output, the demand curve will have the usual negative slope. In contrast, firm A will agree to reduce its own output if its marginal profit loss is compensated by a payment from B. Firm A possesses a supply curve for the reduction of its output which is exactly its own marginal foregone profit curve. If A's profits increase at a decreasing rate, its supply curve will have a positive slope.

The intersection of supply and demand curves[8] defines the optimal level of reduction in output from $\hat{Q}$, firm A's output in the absence of communications between the firms.[9] The intersection does not necessarily define the optimum from society's view; society may prefer a different intersection, for the demand curve incorporates B's adjusted risk attitude, an attitude which may not coincide with society's.

We utilize a neoclassical analysis to determine the supply curve, the demand curve and their intersection, the latter a point which possesses the characteristic that all possible gains from trade have been exhausted. In short, we treat both firms as price takers in the bribery market.

[7]Were A's production to create a positive externality, B would be interested in obtaining an expansion of A's output; the formal analysis would be the same.

[8]Ken-Ichi Inada and Koyoshi Kuga have shown the conditions under which there is no intersection, or no unique intersection

[9]$\hat{Q}$ is defined by $\Pi_Q = 0$, $\Pi_{QQ} < 0$, since A is a riskless profit maximizer who ignores the impact Q has on B.

Following the tradition established by Coase, we assume that transactions between the firms are costless. Barring obtuseness by the managers, bargaining will continue so long as there are gains from trade to be realized. We determine the reduction in Q as a result of bargaining an allocational issue. The actual division of the gains is a distributional question which does not concern us, although it may be of interest in its own right.

Firm A's supply function of Q reduction is unaffected by the uncertain impact its output has on B. From A's view, all relevant parameters are known in advance of the production decision, including the bribery payments it receives. Thus, A maximizes its augmented profit function $\Pi(Q) + P(\hat{Q} - Q)$ where P is the per unit bribery payment and $(\hat{Q} - Q)$ is the level of output reduction.

Firm B continues to confront uncertainty. What it has done in the bribery market is buy a certain reduction in Q and an expected reduction in damages. Firm B maximizes its expected utility from post-bribery profits, found by a Taylor series expansion about $\bar{\pi}^*$, $\{P(\hat{Q} - Q)\}^*$ (the latter term is the optimal bribe).

$$(5)\quad E[u(\pi)] = u[\bar{\pi} - P(\hat{Q} - Q)] + (1/2)u''[\bar{\pi}^* - \{P(\hat{Q} - Q)\}^*]v$$

Firm A's supply function in the bribery market is

$$(6a)\qquad \Pi_Q - P = 0$$

Marginal profits from increased output are balanced against marginal gain from reduced output. For B:

$$(6b)\qquad P + \bar{\pi}_Q = (1/2)r[\cdot]v_Q$$

$$(6c)\qquad \bar{\pi}_q = (1/2)r[\cdot]v_q$$

The latter equation is the standard equilibrium condition: set marginal profits from q equal to the adjusted risk attitude. The former equation says that B's adjusted risk attitude should equal the net marginal profit from a reduction in Q (composed of two parts: the cost of buying the reduction and the savings as a result of the purchase). Firm B's demand curve is (6b) given that (6c) holds.[10]

Combining equations (6) gives the output levels in a world where pollution is permissible and firms may bargain costlessly:

$$(7a)\qquad \Pi_Q + \bar{\pi}_Q = (1/2)r[\cdot]v_Q$$

$$(7b)\qquad \bar{\pi}_q = (1/2)r[\cdot]v_q$$

The effect of imposing legal liability on B is to impose B's adjusted risk attitude on the equilibrium. While these results are similar to those in a merger controlled by B, $r[\cdot]$ is evaluated at a different wealth level. Does this effect the output levels?

THEOREM: *The Coase Theorem is not valid in an uncertain world if the legally liable firm is controlled by a dominant shareholder or an autonomous manager*

PROOF.

Assume (4) and (7) are identical and define $\{Q^*, q^*\}$. All objective values $(\bar{\pi}_q, \bar{\pi}_Q, \Pi_Q, \bar{\pi}^*, v_Q, v_q)$ are identical in both sets of equations. Thus, $r[\bar{\pi}^* - \{P(Q - Q)\}^*] = r(\Pi^* + \bar{\pi}^*)$; but, $\Pi^* \geq 0$ while $-\{P\cdot(Q - Q)\}^* < 0$. Since the risk function is monotonic in wealth, $r[\cdot] \neq r(\cdot)$ and (7) cannot define $\{Q^*, q^*\}$.[11] Of course, if the Arrow-Lind assumptions stated in Section I hold, the Coase Theorem is valid under uncertainty.[12]

[10]Demand is a function of $r[\cdot]$ and, therefore, of B's wealth level A fortiori, demand is dependent on the bargaining process chosen; the total gains from trade are not independent of their distribution.

[11]Two brief comments are in order 1) It is not clear that either (4) or (7) define a first best Pareto optimal allocation since we have precluded the possibility of risk sharing 2) In the event of a merger one would expect that B would compensate A When compensation is paid (4) and (7) may define the same allocation by coincidence

[12]An alternative approach to the stock market, proposed by Franco Modigliani and Merton Miller, assigns firms to risk classes, Jan Mossin has shown that a firm's risk class is $-C\sum_k \sigma_{jk}$, where C is "the same for all companies and can be given an interpretation as market risk aversion" (p. 753). The term $\sum_k \sigma_{jk}$ is the sum of the covariances of profit of company j with all other companies, including itself (i.e., its variance). If the firm's manager takes his risk attitude from the market ($r = -C\sum_k \sigma_{jk}$), the optima defined by equa-

Within the bribery market, the position of B's demand curve is affected by its adjusted risk attitude. If this is zero (certainty equivalence) the supply-demand intersection occurs at a particular price-reduction combination. In contrast, a risk-averse firm will obtain more reduction at a higher price if v_Q is positive, but less reduction at a lower price if v_Q is negative. The reason for this latter case is that extra Q lessens the variance of profits, an event which the risk-averse firm finds attractive. Correspondingly, if $v_Q > 0$, a risk preferrer will demand a smaller reduction and offer a lower per unit bribe than a risk-neutral firm.[13] The risk preferrer is a "tougher bargainer" because it perceives itself as having less to gain from trade— in fact, if B is a sufficiently strong preferrer of risk, there will be no gains from trade available to anyone.[14] Notice that if there are potential gains from trade available, they will generally be divided between the firms Both firms have market power the power to block an agreement is the power to obtain part of the gains. There is absolutely no validity to the naive view expressed by James Marchand and Keith Russell that because B is (effectively) liable, A obtains all the gains from trade.

III. Pollution not Permissible without Compensation

Suppose the legal system permits firm A to befoul the environment to some limit at no penalty but requires that a firm harmed by excessive pollution be fully compensated for its lost profits. If the legal limit is effective—if A exceeds the limit in the presence of penalties—and if B's profits rise with its own output *ceteris paribus*, then A's legally mandated damage payments to B are positively related to q. Firm A has dual incentives: to lower its own output and to persuade B to lessen its production, both events will improve A's profits.

While reduction of Q is an internal matter, reduction of q requires the cooperation of B. Firm B is always willing to curtail its production if it is amply rewarded; its minimal supply price is its relinquished marginal profits. If B's profits increase at a decreasing rate with additional q, B's supply of reduced output will have a positive slope Firm A will demand an output reduction of q so long as it can buy the reduction for no more than its maximal demand price. This is the marginal decrease in its legally mandated "excessive pollution" payment; therefore, A's demand curve is a marginal benefit curve. It will have a negative slope if, as q is curtailed, there is a greater profit reduction for firm B at lower (more acceptable) pollution levels than at higher ones.

Once again, the intersection of the supply and demand curves in the bribery market will define the optimum from the viewpoint of the involved parties. We utilize a neoclassical pricing approach to define the intersection and to distribute the gains from trade, treating the firms as price takers in the bribery market. Reduction occurs from the level $\hat{q}$, the amount of output produced by B in the absence of interfirm negotiation but in the presence of the legally mandated payments.

Firm A's profit is given by

$$\Pi(Q) - [\bar{\pi}(M,q) - \pi(Q,q)] - p(\hat{q} - q) \tag{8}$$

tions (4) and (7) are identical given the present formulation of the problem. That society's risk attitude should be the same as the private risk attitude for this risk class has been shown by Agnar Sandmo However, a slight reformulation of our problem will cause the optima of (4) and (7) to diverge Write A's profits as $\Pi(Q) + \epsilon$, $\epsilon \sim (0, \sigma^2)$, so that $\epsilon_q = 0 = \epsilon_Q$. (This construction of A's profit function will leave its first-order maximization conditions unaffected.) Let the covariance of A's and B's profits be zero with respect to all other firms in the market Then the merged firm has as its risk attitude $-C(\sigma^2 + v + COV(\pi, \Pi))$. When pollution is legally permissible, the market assigns it to the risk class $-C(v + COV(\pi, \Pi))$. Firm B does not take cognizance of the system uncertainty which confronts A; thus, B has a different risk class than does the merged firm and equations (4) and (7) are not identical

[13] Ira Horowitz, p 367, reaches a similar conclusion for a related problem.

[14] If B is an extreme risk preferrer, the supply and demand curves intersect to the left of the vertical axis. Firm B demands an increase in Q.

where $p(\hat{q} - q)$ is the bribe paid; its level is determined by costless negotiation prior to production. Firm A's profit from its own output ($\Pi(Q)$) is also known a priori. The legally mandated damage payment to B – the bracketed term – is known only after the fact. Thus, A confronts uncertainty. Firm B is guaranteed the difference between (a) the expected level of its own profits when $Q = M (M \geq 0)$, the output creating the mandated level of expected externalities, and (b) the actual level of its profits. Notice that when the state of nature retards pollution, $\pi(Q,q)$ rises and damage payments fall. Firm A, not B, benefits from a favorable state of nature; A, of course, will maximize its expected utility from net profits.

Firm B's guaranteed profits are $\pi(Q,q) + [\bar{\pi}(M,q) - \pi(Q,q)] + p(\hat{q} - q) = \bar{\pi}(M,q) + p(\hat{q} - q)$. B no longer confronts uncertainty.[15] In effect, the legal system causes A to insure B against risk. Firm B's supply curve is

$$\bar{\pi}_{q|M} - p = 0 \tag{9a}$$

where $\bar{\pi}_{q|M} \equiv \partial\bar{\pi}(M,q)/\partial q$. A's first-order conditions, found after utilizing a Taylor series to expand the expected utility of (8) about Π^*, $\bar{\pi}^*$, $\bar{\pi}^*(M,q)$, $\{p(\hat{q} - q)\}^*$, are

$$\Pi_Q + \bar{\pi}_Q = (1/2)R[\cdot]v_Q \tag{9b}$$

$$p = (1/2)R[\cdot]v_q + (\bar{\pi}_{q|M} - \bar{\pi}_q) \tag{9c}$$

Firm B's supply of reduced output (9a) is defined by the condition that marginal bribery gain equal marginal profit loss. Equations (9b) and (9c) define A's demand curve. The former says that marginal profits from Q, net of external damage to B, should equal A's adjusted risk attitude. The latter sets the marginal bribery cost p equal to the adjusted risk attitude plus the marginal legal required expected damage payment. Combining equations (9) yields a set which characterizes the equilibrium position:

$$\Pi_Q + \bar{\pi}_Q = (1/2)R[\cdot]v_Q \tag{10a}$$

$$\bar{\pi}_q = (1/2)R[\cdot]v_q \tag{10b}$$

These are identical to the Pareto optimal conditions (4) only under the same restrictions given for equations (7) in Section II.

We have shown that the assignment of liability for externalities may be of consequence. The final allocation of resources is dependent upon liability rules when there is uncertainty as to the state of nature and when indivisibilities preclude the complete sharing of risk.

IV. Social Intervention and Pareto Optimality

Suppose costless negotiation does not lead to the optimum defined by equations (4). In addition to the reasons stated above, "... if one accepts the proposition that the state is more than a collection of individuals and has an existence and interests apart from those of its individual members, then it follows that government policy need not reflect individual preferences" (Arrow-Lind, p. 365). Thus, the government may wish to intervene even if equations (7) and/or (10) define $\{Q^*, q^*\}$. Second best questions aside, can the government improve (in its view) the allocation of resources? The answer is yes; per unit and/or lump sum taxes (or subsidies) may be utilized.

To economize on space, we investigate only the case of permissible pollution. We start by stating a set of simplifying assumptions. They are not crucial to the analysis. Let $S(\cdot) = 0$, $r(\cdot) > 0$, and the Principle of Increasing Uncertainty hold (i.e., $v_t > 0$). The optimum is defined as $\Pi_Q + \bar{\pi}_Q = 0 = \bar{\pi}_q$. Let there be a tax T per unit of output Q levied on firm A and another tax t per unit of output q levied on firm B. Bargaining still occurs. Profits for firms A and B, respectively, are

[15] B could refuse to bargain with A, produce the (then optimal) output level $\hat{q}$, and sue A for damages done: damages which are related to Q, $\hat{q}$, and the state of nature known to have prevailed when production occurred. Firm B rejects this avenue because its total profits $[\bar{\pi}(M, \hat{q})]$ from the lawsuit would be less than those it can obtain by bargaining. Notice that if $Q \leq M$ in the absence of negotiations, or if $\bar{\pi}_Q > 0$, we belong in Section II.

$$(11a) \quad \Pi(Q) + P(\hat{Q} - Q) - TQ$$

$$(11b) \quad \pi(Q,q) - P(\hat{Q} - Q) - tq$$

Firm B continues to confront uncertainty; thus, it continues to maximize its expected utility of profits. First-order conditions are now:

$$(12a) \quad \Pi_Q - P - T = 0$$

$$(12b) \quad \bar{\pi}_Q + P - (1/2)r\{\cdot\}v_Q = 0$$

$$(12c) \quad \bar{\pi}_q - t - (1/2)r\{\cdot\}v_q = 0$$

where $\{\cdot\} \equiv \{\bar{\pi}^* - [P(\hat{Q} - Q)]^* - (tq)^*\}$. Note that T is not independent of P. While we treat P as parametric for ease of presentation, it is in practice determined by negotiation. Thus, when establishing tax/subsidy levels, the government must consider its own impact upon the bargaining process. Equations (12) combine to form

$$(13a) \quad \Pi_Q + \bar{\pi}_Q = (1/2)r\{\cdot\}v_Q + T$$

$$(13b) \quad \bar{\pi}_q = (1/2)r\{\cdot\}v_q + t$$

The optimal tax levels are $T = -(1/2)r\{\cdot\}v_Q < 0$ and $t = -(1/2)r\{\cdot\}v_q < 0$; the optimal taxes are subsidies because firm B is more risk averse than society.[16]

There are three points of interest here. First, apart from questions of income distribution, $-TQ$ could be replace by $+T(\hat{Q} - Q)$. The effect on the shadow price of the marginal unit of Q is what matters and it would be unaltered. Second, $-TQ$ (or $+T(\hat{Q} - Q)$) could appear in (11b) instead of (11a); equations (13) would be of the same form although the value of $r\{\cdot\}$ would in general change. Third, neither T nor t can be set equal to the right-hand side of equations (7) because nonzero taxes have an effect upon the wealth level at which $r\{\cdot\}$ is evaluated. This final point can be seen by taking the total derivative of (12b) and (12c) and rearranging terms. We obtain

$$(14a) \quad \frac{\partial Q}{\partial t} = \frac{qr'(v_q L_{Qq} - v_Q L_{qq})}{2D} - \frac{L_{Qq}}{D}$$

$$(14b) \quad \frac{\partial q}{\partial t} = \frac{qr'(v_Q L_{Qq} - v_q L_{QQ})}{2D} + \frac{L_{QQ}}{D}$$

where r' is the rate of change of the Pratt-Arrow measure of absolute risk aversion due to a change in wealth; $L_{qq} < 0$, $L_{QQ} < 0$, $D \equiv L_{QQ}L_{qq} - L_{Qq}^2 > 0$, all from second-order conditions; and P is parametric.

Equations (14) may be expressed more simply as:

$$(15a) \quad \frac{\partial Q}{\partial t} = q\frac{\partial Q}{\partial f} - \frac{L_{Qq}}{D}$$

$$(15b) \quad \frac{\partial q}{\partial t} = q\frac{\partial q}{\partial f} + \frac{L_{QQ}}{D}$$

where $(\partial Q/\partial f)$ and $(\partial q/\partial f)$ are the effect on output levels of a change in fixed costs.[17] Their signs are dependent upon whether the firm's risk attitude is increasing, invariant, or decreasing in wealth ($r' \gtreqless 0$) and upon the value of the parenthetical term in (14). If $L_{Qq} \geq 0$, our earlier assumptions with the second-order conditions guarantee that $(\partial Q/\partial f)$ and $(\partial q/\partial f)$ are of the same sign[18] as r'. However, if $L_{Qq} < 0$, we cannot in general determine the effect a change in wealth has upon output. Now, from equations (15), we see that a change in the per unit subsidy (from, say, zero) has a dual impact upon output levels. The first is the income effect: q times a pure wealth effect; the second is a substitution effect which is negative for $(\partial q/\partial t)$ and of uncertain sign for $(\partial Q/\partial t)$. Note that a lump sum tax has a pure wealth effect on output levels since it is equivalent to a change in fixed costs.

Consider now an impediment to bargaining which prevents any interfirm negotiations. Then $P = 0$ in equations (11) and (12a), while (12b) does not "exist" because

[16]There would be a positive tax if society were the more risk averse. Note that if $v_Q > 0$ and $v_q < 0$, society would be in the peculiar position of subsidizing the polluting output and taxing the output of q.

[17]They are obtained by explicitly considering fixed costs f, by writing profits as $\pi(Q, q) - f$, and totally differentiating the first-order conditions. The effect of f on the first-order conditions occurs only in the wealth level at which $r[\cdot]$ is evaluated.

[18]$L_{Qq} = 0$ is not guaranteed by an additively separable profit function

Q is not a choice variable for firm B. The government can still create an optimal allocation of resources by setting $T = -\bar{\pi}_Q$ and $t = -(1/2)r\{\cdot\}v_q, \{\cdot\} \equiv \{\bar{\pi}^* - (tq)^*\}$. The first point mentioned in association with equations (13) remains valid, as does the third. The second no longer holds due to the lack of communication. This communicative absence is corrected by the tax T. The subsidy t is used to rectify misallocations caused by the deviation of B's adjusted risk attitude from society's attitude.

In the case of (bribery) market failure the government can, in principle, intervene to create a Pareto optimal allocation of resources. However, the wealth effect (absent with certainty equivalence) compels the government to know firm B's risk function, not just its value at a (nonoptimal) set of output levels defined by equation (7).

V. Conclusion

When one firm's productive process imposes an externality, positive or negative, upon another firm, those enterprises have cause to attempt to interact in a private bribery market in order to improve both of their profit levels. The nature of the bribery market is determined by the legal system. If A is liable for damages, it will demand of B a reduction in B's output (and thereby achieve a cutback in its damage payments). Firm B will supply an output curtailment so long as its marginal profits foregone are covered by A. The intersection of supply and demand in the bribery market defines the equilibrium position from which no gains from trade remain; thus, at the close of bargaining no Pareto-relevant externalities exist, although there are (in general) externalities present. When A is not liable for damages, the same type of analysis and the same conclusions apply, the difference is that it is B which demands an output reduction from A. These are the standard externality results from which many authors have concluded that the socially optimal allocation of resources occurs as an outcome of the costless bargaining process. The allocation is, they say, independent of the assignment of liability.

When the externality is randomly distributed, uncertainty confronts the firm which is effectively liable. This firm must then incorporate its own attitude toward risk along with the distribution of the externality into its decision rule. We have utilized a Taylor series expansion to embody these facts of economic life into the enterprise's optimization process, a maximization which occurs across the expected utility of profits.

In an uncertain world, liability rules may determine resource allocation. Costless bargaining is not sufficient to guarantee that the Coase Theorem holds; it is also necessary that risk may be shared (say through a stock market) and that there be no indivisibilities. Without the possibility of sharing risk the involved firms will have the same attitude towards risk only by accident. Indivisibilities—an autonomous manager or a dominant shareholder—also preclude a complete sharing of risk even in the presence of a stock market.

When these stronger assumptions do not obtain, the risk attitude of the manager of the firm made responsible for damages is embodied in the equilibrium conditions which derive from bargaining. Resource allocation is affected by legal liability as well as by the bargaining skills of the involved firms. However, the government, by use of a tax/subsidy scheme, can intervene to create a Pareto optimal level of outputs. The attractive results of the Coase Theorem (as usually stated) are a special case in an uncertain world.

APPENDIX

A

Here we examine the output effect upon both firms of a change in the distribution of π, concentrating upon the legal rule of permissible pollution. We investigate three types of change: a linear shift of expected profits, holding all moments about the mean constant; a spreading of the distribution about a constant mean, summarized as

a multiplicative shift of the variance; and a proportional shift of all moments.

The three cases may be stated as

$$A: \pi \sim (\alpha + \bar{\pi}, v)$$
$$B: \pi \sim (\bar{\pi}, \beta v)$$
$$C: \pi \sim (\lambda \bar{\pi}, \lambda^2 v)$$

At this time we explicitly introduce fixed costs by writing profits as $\pi(Q,q) - f$, ($f \geq 0$), in order to isolate an income effect. Of course, f is functionally equivalent to a lump sum tax.

Case A. The procedure is to substitute $\alpha + \bar{\pi}$ for $\bar{\pi}$ in equations (6). Consider a marginal change in α from its initial level such that the expected utility of profits remains maximized. This requires that we take the total derivative and evaluate it at $\alpha = 0$. Manipulation shows

$$\text{(A1)} \qquad \partial X/\partial \alpha = -\partial X/\partial f$$

where X designates Q or q. A marginal increase in expected profits has the same effect as a marginal decrease in fixed costs. There is a pure income effect upon both output levels.

To obtain (A1) we assumed that the firm incurred no cost to create a change in α. Since the firm might invest in pollution abatement equipment, it seems worth pointing out that such a situation would add another term to the right-hand side of (A1): $+ (\partial X/\partial f)(\partial f/\partial \alpha)$. Because the firm would voluntarily invest in abatement equipment only if $(\partial \alpha/\partial f) > 1$, we can state that the sign of $(\partial X/\partial \alpha)$ is the negative of the sign of $(\partial X/\partial f)$ whether a change in α is endogenous or exogenous.

Case B: The procedure is the same as in Case A, with $(\partial f/\partial \beta) = 0$ and the total derivative evaluated at $\beta = 1$, the original variance level. Thus,

$$\text{(A2)} \qquad \frac{\partial X}{\partial \beta} = -\left[\frac{\partial X}{\partial f}\right]\left[\frac{r}{r'}\right]$$

We now obtain a weighted income effect. The weight is the ratio of firm B's measure of absolute risk aversion to its rate of change. An increasingly risk-averse firm ($r, r' > 0$) will respond to spreading distribution in the same manner as to an increase in expected profit. Conversely, a decreasingly risk-averse firm ($r' < 0 < r$) will react oppositely to an increased variance than to an increased mean.

Case C: The procedure is as above, with evaluation occurring at $\lambda = 1$. The output effect of a proportional change in all moments is

$$\text{(A3)} \qquad \frac{\partial Q}{\partial \lambda} = \left[\frac{\partial Q}{\partial \beta} + \bar{\pi}\left(\frac{\partial Q}{\partial \alpha}\right)\right] + \frac{PL_{qq}}{D}$$

$$\text{(A4)} \qquad \frac{\partial q}{\partial \lambda} = \left[\frac{\partial q}{\partial \beta} + \bar{\pi}\left(\frac{\partial q}{\partial \alpha}\right)\right] - \frac{PL_{Qq}}{D}$$

There is an income effect and a substitution effect. The bracketed term is similar to the preceding case, although determination of the sign is less obvious if r' and r are not of the same sign. The substitution effect is of determinable sign only for (A3). Since $D > 0$ and $L_{qq} < 0$ by the second-order conditions, and since the marginal bribe (P) is positive, the substitution effect for $(\partial Q/\partial \lambda)$ is negative. The substitution effect for $(\partial q/\partial \lambda)$ is of the opposite sign from the (unknown) sign of L_{Qq}.

B

In the body of this paper we considered what is essentially a polar case; there is no mechanism for the sharing of risk. In this polar case the Coase Theorem does not hold strictly. It was argued that in the opposite polar case, in which there is perfect sharing of risk, the Coase theorem holds. This part of the appendix considers briefly an intermediate case in which the parties may share risk. Assume that B is liable for damages and as a consequence has uncertain profits. The essential trade is for B to transfer to A a portion of the deviation of his profit from its expected value and for B to compensate A for the service A renders. Since A absorbs some risk his expected

utility depends in part on q and there is the possibility of a market existing for a change in the level of q. Allowing for these possibilities we write B's augmented profits as

$$\bar{\pi}(Q,q) + \delta[\pi(Q,q) - \bar{\pi}(Q,q)] - P(\hat{Q} - Q) + p(\hat{q} - q) - b(1 - \delta) \tag{A5}$$

where δ is the share of B's profit deviation retained by B, and b is the unit price B pays A for taking a share. Firm B's expected utility is

$$u[\bar{\pi}(Q,q) - P(\hat{Q} - Q) + p(\hat{q} - q) - b(1 - \delta)] + (1/2)u''(\cdot)\delta^2 v \tag{A6}$$

This is maximized where:

$$\begin{aligned} \bar{\pi}_Q + P &= r(\delta)^2 v_Q \\ \bar{\pi}_q - p &= r(\delta)^2 v_q \\ b &= 2r\delta v \end{aligned} \tag{A7}$$

By means of a similar exercise A's expected utility is

$$U[\Pi(Q) + P(\hat{Q} - Q) - p(\hat{q} - q) + b(1 - \delta)] + (1/2)U''(\cdot)(1 - \delta)^2 v \tag{A8}$$

which is maximized when

$$\begin{aligned} \Pi_Q - P &= R(1 - \delta)^2 v_Q \\ p &= R(1 - \delta)^2 v_q \\ b &= 2R(1 - \delta)v \end{aligned} \tag{A9}$$

Letting each market clear simultaneously leaves

$$\begin{aligned} \Pi_Q + \bar{\pi}_Q &= (Rr/(R + r))v_Q \\ \bar{\pi}_q &= (Rr/(R + r))v_q \end{aligned} \tag{A10}$$

When A is liable a similar set of conditions is found, but they will not in general imply the same allocation since the risk attitudes will not be evaluated at the same income levels.

REFERENCES

Kenneth Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1971.

______ **and R. Lind,** "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

R. Coase, "The Problem of Social Cost," *J. Law Econ*, Oct. 1960, *3*, 1–44.

O. Davis and A. Whinston, "Some Notes on Equating Private and Social Cost," *Southern Econ. J.*, Oct. 1965, *32*, 113–26.

Ira Horowitz, *Decision Making and the Theory of the Firm*, New York 1970.

K. Inada and K. Kuga, "Limitations of the 'Coase Theorem' on Liability Rules," *J. Econ. Theory*, Dec. 1973, *6*, 606–13.

H. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

O. Loistl, "The Erroneous Approximation of Expected Utility by Means of a Taylor's Series Expansion: Analytic and Computational Results," *Amer. Econ Rev.*, Dec. 1976, *66*, 904–10.

J. Marchand and K. Russell, "Externalities, Liability, Separability, and Resource Allocation: Reply," *Amer. Econ. Rev.*, Sept. 1975, *65*, 730–32.

F. Modigliani and M. Miller, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261–97.

J. Mossin, "Security Pricing and Investment Criteria in Competitive Markets," *Amer. Econ. Rev.*, Dec. 1969, *59*, 749–56.

C. Plott, "Externalities and Corrective Taxes," *Economica*, Feb. 1966, *33*, 84–87.

J. Pratt, "Risk Aversion in the Large and in the Small," *Econometrica*, Jan.–Apr. 1964, *32*, 122–36.

A. Sandmo, "Discount Rates for Public Investment Under Uncertainty," *Int. Econ.* Rev., June 1972, *13*, 287–303.

Production, Efficiency, and Welfare in the Natural Gas Transmission Industry

By Jeffrey L. Callen*

In a recent conference, Leland Johnson criticized the asymmetry between the voluminous theoretical literature and the absence of empirical work on the Averch-Johnson-Wellisz (A-J-W) hypothesis. This criticism has engendered a body of empirical research that is characterized by other imbalances. With one notable exception,[1] the empirical A-J-W literature concentrates on the electrical generating industry disregarding the potential benefits from intellectual diversification.[2] Also, much effort is expended testing for the overcapitalization phenomenon, but the attendant welfare implications are virtually ignored. This paper tries to redress some of these imbalances by (i) investigating the A-J-W hypothesis in the *U.S.* interstate natural gas transmission industry, and (ii) analyzing the social welfare impact of rate of return regulation on this industry.

In what follows, Section I formulates the optimization models which are used to simulate the input-output decision of a natural gas transmission company. Of the four models which are developed in this section, two are independent of the regulatory environment, while the other two are constrained by it. A brief description of the constraint employed by the Federal Power Commission (*FPC*) to regulate interstate transmission companies precedes the formulation of the constrained models. Also this section discusses the input distortions which are potential consequents of rate of return regulation. Section II compares the simulated solutions with data from a comprehensive sample of natural gas transmission companies. The model which predicts the best is presumed to reflect the underlying behavior of the industry. Section III analyzes the social welfare implications of regulating the industry by comparing the benefits of rate of return regulation to those obtainable from marginal-cost pricing.

I. The Models

A. *Some Preliminary Assumptions*

Transmission revenues are derived from three sources: the transmission and sale of natural gas to other transmission companies and retail distributors, commonly known as sales for resale Q_1; the transmission and direct sale of gas to large industrial corporations, commonly known as main line industrial sales Q_2; and the nonsale transmission of gas for other pipelines Q_3 net of own gas transmitted by other pipelines Q_4. It is assumed that Q_2, Q_3, and Q_4 are proportional to Q_1. This assumption is necessary for computational simplicity, and reasonable because transmission activities other than sales for resale are minor or negligible for most large transmission companies.[3] Other assumptions concerning the models follow. The demands for Q_1 and Q_2 are governed by constant elasticity demand

*Assistant professor of finance and business economics, McMaster University This paper is adapted from my doctoral dissertation submitted to the Faculty of Management Studies, University of Toronto and written under the guidance of G. David Quirin, Basil Kalymon, Frank Mathewson, and Jack Sawyer I am also indebeted to Varouj Aivazian, George Borts, Danny Frances, Stan Laiken, and Herbert Mohring for their comments on earlier drafts

[1]Paul MacAvoy and Roger Noll also, study the natural gas transmission industry but their methodology differs fundamentally from mine. They are concerned with the impact of regulation on prices rather than inputs and output They also disregard the social welfare ramifications of pipeline regulation

[2]Foremost among this expanding literature are the studies by Leon Courville, Thomas Cowing, Paul Hayashi and John Trapani, H. Craig Peterson, and Robert Spann.

[3]This is especially true of those companies which comprise my sample. See fn. 11 below and *FPC Statistics* 1965.

functions $P_1(Q_1)$ and $P_2(Q_2)$. The prices of Q_3 and Q_4, P_3 and P_4, are constant. Transmission costs, both fixed and variable but excluding the cost of the gas, are proportional to either horsepower capacity H or line-pipe capacity K. The wholesale purchase price of a unit of gas ϕ is constant.

B. *The Unconstrained Models*

In the absence of a regulatory constraint, either cost minimization or profit maximization are assumed to determine a transmission company's behavior. The firm's behavior is further limited by the technology relationship developed in Appendix A. Specifically, the profit-maximizing (*PM*) model can be represented by the following: Maximize (1) with respect to Q_1, H, K:

$$(1)\quad V = (1-\tau)[P_1(Q_1)Q_1 + P_2(Q_2)Q_2 + P_3Q_3 - P_4Q_4 - \phi(Q_1+Q_2)] - [(1-\tau)W_v + (r-\tau d_H)W_F]H - [(1-\tau)P_v + (r-\tau d_K)P_F]K$$

subject to

$$(2)\quad Q_1 = AH^{.27}K^{.9}$$

where

P_F = fixed costs per unit of line-pipe capacity
W_F = fixed costs per unit of horsepower capacity
P_v = variable costs per unit of line-pipe capacity
W_v = variable costs per unit of horsepower capacity
τ = corporate tax rate
r = firm's (weighted average) cost of capital
d_H = depreciation rate for horsepower related equipment
d_K = depreciation rate for line pipe
A = scale constant

The cost-minimizing (*CM*) model solves for the same input ratio (H/K) as the *PM* solution but an indeterminate output level.

C. *The Regulatory Constraint*

The revenues earned by an interstate natural gas transmission company on Q_1, its sales for resale also called jurisdictional sales, are regulated by the *FPC* through the Atlantic Seaboard cost allocation formula. The initial step in applying this formula necessitates estimating the cost of service—operating expenses, taxes, depreciation, and a "fair" return to shareholders—on the basis of test year data. The components of the cost of service are then allocated to either a demand or commodity cost classification. Theoretically, the demand classification is comprised of those costs incurred providing fixed pipeline capacity. The commodity classification, on the other hand, should include both the cost of the gas and the variable costs of transmitting it to the customer. In practice, the Atlantic Seaboard formula splits most costs evenly between the two classifications with some important exceptions. The cost of the gas and pipeline produced gas expenses are allocated entirely to the commodity classification. Most, but not all, compressor and production expenses are allocated to the commodity classification. Demand charges levied by one transmission company on another in interstate sales are included in the buyer's demand classification.

Revenues derived from other than transmission-sales activities are netted against the cost of service. Nonsales transmission and storage revenues are credited wholly to the commodity classification. Other revenues, which are usually derived from the sale of natural gas by-products, are credited equally to each classification.

The next step in applying the Atlantic Seaboard formula involves allocating costs between jurisdictional and nonjurisdictional markets. The commodity classification is weighted by the ratio of jurisdictional to total annual sales. Demand is weighted by the ratio of jurisdictional to total "firm" sales during a three-day sustained peak period. The sum of the two is the cost of service attributable to the jurisdictional

market which serves as an upper bound on the revenues the company is allowed to earn in the jurisdictional market.

D. *The Constrained Models*

Two possible objectives are postulated for the constrained models, profit maximization and revenue maximization. In addition to the technology relationship from Appendix A, the firm's input-output decisions are mediated by a regulatory constraint patterned after the Atlantic Seaboard formula shown below as constraint (5).[4] Formally, the constrained revenue-maximizing (CRV) model can be represented by:[5] Maximize (3) with respect to Q_1, H, K:

$$(3) \quad V^1 = P_1(Q_1)Q_1 + P_2(Q_2)Q_2 + P_3Q_3 - P_4Q_4$$

subject to

$$(4) \quad Q_1 = AH^{.27}K^{.9}$$

$$(5) \quad P_1(Q_1)Q_1 \leq \frac{Q_{1B}}{Q_{1B} + Q_{2B}} \cdot [(1/2)(P_vK + W_{1v}H + T) + (1/2)(P_FK + W_FH + T^1)(1 - \delta)s + (1/2)(d_KP_FK + d_HW_FH + d^1T^1) + (1/2)\tau^* + (1/2)P_4Q_4$$
$$+ M\phi(Q_1 + Q_2) - (1/2)\,(\text{miscellaneous revenues})]$$
$$+ \frac{Q_1}{Q_1 + Q_2}[(1/2)(P_vK + W_{1v}H + T) + (1/2)(P_FK + W_FH + T^1)(1 - \delta)s + (1/2)(d_KP_FK + d_HW_FH + d^1T^1) + (1/2)\tau^* + (1/2)P_4Q_4 + (1 - M)\phi(Q_1 + Q_2) + W_{2v}H - P_3Q_3 - (1/2)(\text{miscellaneous revenues}) + (\text{gas production and gathering expenses})]^6$$

where

Q_{1B} = sales for resale peak load demand
Q_{2B} = main line industrial sales peak load demand
W_{1v} = variable costs per unit of horsepower capacity allocated to both demand and commodity classifications
W_{2v} = variable costs per unit of horsepower capacity allocated entirely to the commodity classification
M = proportion of production expenses allocated to the demand classification
T = variable costs unrelated to K or H
T^1 = fixed costs unrelated to K or H
τ^* = income and property taxes[7]
s = fair rate of return
δ = accumulated depreciation rate
d^1 = average depreciation rate for assets unrelated to K or H

The constrained profit-maximizing (CPM)

[4]Constraint (5) is a proxy for the Atlantic Seaboard formula The first set of square brackets on the right-hand side of the inequality contains the demand costs. These are weighted by the ratio of jurisdictional to total firm sales during the peak period of the test year. The second set of square brackets contains the commodity costs which are weighted by jurisdictional to total sales during the entire test year Therefore, those revenues which the firm may earn in the jurisdictional market are restricted by constraint (5) not to exceed *ex ante* the sum of demand and commodity costs attributable to the jurisdictional market in the test year. Although uncertainty about sales and regulatory lag may cause the constraint to be violated *ex post*, it is assumed that these effects are not endogenized by the firm during the pipeline planning stage and any excess profits earned thereby are treated as windfall gains.

[5]The constrained models are solved by a dual iteration-linearization technique described by the author, pp. 102–07.

[6]Miscellaneous revenues, and gas production and gathering expenses are treated as constants. The allowance for working capital and interest during construction are not part of the constraint formulation since they are trivial amounts and the relevant data are not available.

[7]The level of taxes τ^* is treated as a function of the other parameters and variables, $\tau^* = \tau[P_1(Q_1)Q_1 + P_2(Q_2)Q_2 + P_3Q_3 - P_4Q_4 - \phi(Q_1 + Q_2) - W_VH - P_VK - (d_KP_FK + d_HW_FH) - T$ - other deductions on corporate income taxes] + other state and local (property) taxes.

model is identical to the CRV model except that V^1 (equation (3)) is replaced by V (equation (1)).

E. *Regulatory Input Biases*

The principal variable input into the transmission process, compressor fuel, cannot be forecast accurately.[8] Therefore the models simulate, in addition to output, only the two fixed inputs H and K. Nevertheless, the input biases appear in the tradeoff between the H horsepower capacity and K line-pipe capacity variables. Intuition and the A-J-W literature suggest that, since H has a large variable cost component and line-pipe expenses are trivial, the CPM model's simulated K/H input ratio is always larger than the cost-minimizing ratio. This is not the case, however. The CPM input ratio may prove to be less than the corresponding cost-minimizing solution. To see this, consider the simple case of a firm which deals only in sales for resale for which the CPM model can be represented by: Maximize (6) with respect to K, H:

$$(6) \quad V'' = (1-\tau)x[Q_1(K,H)] - wH - pK$$

subject to

$$(7) \quad (1-\tau)x[Q_1(K,H)] - \overline{w}H - \overline{p}K - G \leq 0$$

where x is the excess of revenues over the cost of the gas, $Q_1(K,H)$ the production function, and w, p, $\overline{w}$, $\overline{p}$, and G are constants.[9] Forming the appropriate Lagrangian, the first-order conditions excluding the constraint are

$$(8) \quad (1-\tau)(1-\lambda)x'Q_{1K} = p - \lambda\overline{p}$$

$$(9) \quad (1-\tau)(1-\lambda)x'Q_{1H} = w - \lambda\overline{w}$$

where λ is the multiplier.[10] Therefore, the marginal rate of technical substitution is

$$(10) \quad \frac{Q_{1K}}{Q_{1H}} = \frac{p}{w} + \frac{\lambda p(\overline{w}/w - \overline{p}/p)}{(w - \lambda\overline{w})}$$

so that the relationship between the CPM and cost-minimizing input ratio depends on the a priori indeterminate signs of both $(\overline{w}/w - \overline{p}/p)$ and $(w - \lambda\overline{w})$. The sign of the former is a function of the relative magnitudes of cost and regulatory parameters which differ from one company to another. The sign of the latter as seen in equation (9) is determined by whether marginal revenue is greater than or less than the marginal cost of the gas at the optimum. The bias in the CRV model's input ratio is also a function of the sign of $(\overline{w}/w - \overline{p}/p)$, but in the majority of cases the CRV model's K/H input ratio is less than that of the cost-minimizing alternative.

II. Comparing the Data with the Simulated Solutions

A. *The Data and the Simulated Solutions*

The 1965 output Q, horsepower capacities, and line-pipe capacities of twenty-eight "major" interstate natural gas transmission companies are listed in the first three columns of Table 1.[11] The corresponding

[8]Natural gas compressor-prime mover units are either reciprocating-gas engine or centrifugal-gas turbine. The latter consume a significantly greater amount of fuel per horsepower generated than the former, for a given horsepower capacity. Therefore, without an inventory of compressor types (for each firm) fuel consumption cannot be estimated. Nor is it reasonable to assume a representative inventory since the proportion of compressor-prime mover types differs dramatically among firms for which the data are available. See James Jensen and Thomas Stauffer, pp. 93-95.

[9]Using the previous notation
$w = (1-\tau)W_V + (r - \tau d_H)W_F$
$p = (1-\tau)P_V + (r - \tau d_K)P_F$
$\overline{w} = (1-\tau)W_V + ((1-\delta)s + (1-\tau)d_H)W_F$
$\overline{p} = (1-\tau)P_V + ((1-\delta)s + (1-\tau)d_K)P_F$
$G = (1-\tau)T + ((1-\delta)s + d^1)T^1$ + gas production and gathering expenses + τ (other deductions on corporate income taxes) − other state and local (property) taxes

[10]E. E. Zajac shows that $0 < \lambda < 1$ if the constraint is binding.

[11]My sample is restricted to major pipeline companies as defined by the *FPC Statistics* 1965, p. viii. The remaining interstate pipeline companies are either distribution companies or have small pipeline systems which cannot be described by a Cobb-Douglas production technology. In addition, four of the thirty-two major companies were excluded because they primarily transport gas owned by their affiliates. The excluded major companies are Chicago District Pipeline, Columbia Gulf Transmission, Florida Gas Transmission, and Humble Gas Transmission.

TABLE 1—1965 ACTUAL AND SIMULATED PROFIT-MAXIMIZING (*PM*) INPUT-OUTPUT SOLUTIONS

	Actual			Simulated *PM*		
Company	*Q* (1)	*H* (2)	*K* (3)	*Q* (4)	*H* (5)	*K* (6)
1. Algonquin Gas Transmission	101	31	141	22	15	32
2. American Louisiana Pipeline	200	142	569	52	28	206
3. Atlantic Seaboard	223	48	239	48	11	66
4. Cities Service Gas	420	197	568	92	39	172
5. Colorado Interstate Gas	332	82	235	53	11	57
6. Consolidated Gas Supply	345	104	370	93	19	142
7. El Paso Natural Gas	1,413	746	2,199	524	154	1,172
8. Kentucky Gas Transmission	97	6	65	18	2	15
9. Manufacturers Light and Heat	209	25	227	48	18	60
10. Michigan Gas Storage	90	16	102	22	4	32
11. Michigan Wisconsin Pipe Line	339	200	632	105	39	281
12. Midwestern Gas Transmission	218	27	265	52	9	75
13. Mississippi River Transmission	206	111	248	55	19	96
14. Natural Gas Pipeline Co. of America	659	525	2,080	151	104	660
15. Northern Natural Gas	507	593	1,650	194	132	889
16. Ohio Fuel Gas	399	33	305	99	13	86
17. Pacific Gas Transmission	194	26	307	51	11	88
18. Panhandle Eastern Pipe Line	600	445	1,256	128	63	406
19. South Texas Natural Gas Gathering	115	5	33	26	2	9
20. Southern Natural Gas	427	277	642	84	60	168
21. Tennessee Gas Transmission	1,055	999	3,233	214	173	931
22. Texas Eastern Transmission	792	902	2,068	203	164	757
23. Texas Gas Transmission	559	276	847	144	62	293
24. Transcontinental Gas Pipe Line	636	647	1,978	198	166	789
25. Transwestern Pipeline	179	64	385	59	23	152
26. Trunkline Gas	314	193	670	91	41	270
27. United Fuel Gas	337	98	187	72	9	68
28. United Gas Pipe Line	1,306	169	1,109	343	85	309

Note: *Q* is measured in billions of cubic feet of natural gas per year, *H* in thousands of horsepower, and *K* in thousands of tons of line pipe. Details concerning the data sources are found in Appendix B.

simulated *PM* solutions for each of these firms are presented in the last three columns of Table 1 and the *CPM* and *CRV* solutions in Table 2. Detailed descriptions of the data base, and some of the variable and parameter estimates utilized in the simulations are found in Appendix B.

B. *Comparing the Input Ratios*

On the input side, the predictive abilities of the models are evaluated by comparing the simulated and actual K/H ratios using the absolute prediction error criterion

$$(11) \qquad \left| \frac{(K/H)_S - (K/H)_A}{(K/H)_A} \right|$$

where A and S stand for the actual and simulated solution, respectively. These prediction errors are found in the first three columns of Table 3. Since the *CM* and *PM* ratios are identical, their prediction errors are listed in the same column. The fourth column of Table 3 gives the best input model for each firm in the sample where the best input model is the one with the smallest absolute error. Of the twenty-eight cases, the *CRV* model predicts the best in ten, the *CPM* model in ten, and the *CM-PM* models in eight. The average absolute prediction errors are 8.81, .73, and .69 for the *CPM*, *CM-PM*, and *CRV* models, respectively.

Additional evidence on the predictive superiority of the *CRV* model can be obtained from the ordinary least squares estimate of b in the equation

$$(12) \qquad (K/H)_A = b(K/H)_S + \epsilon$$

TABLE 2 – SIMULATED CONSTRAINED PROFIT-MAXIMIZING (*CPM*) AND CONSTRAINED REVENUE-MAXIMIZING (*CRV*) INPUT-OUTPUT SOLUTIONS

	CPM			*CRV*		
Company	*Q* (1)	*H* (2)	*K* (3)	*Q* (4)	*H* (5)	*K* (6)
1	100	47	123	100	51	120
2	171	189	437	181	101	562
3	221	31	270	221	40	250
4	316	158	443	328	89	548
5	235	46	190	235	37	204
6	93	6	202	129	30	180
7	1,024	44	3,603	1,319	397	2,462
8	95	3	75	97	10	55
9	129	25	135	129	22	139
10	80	38	69	85	14	99
11	350	266	601	365	142	761
12	222	52	222	223	37	248
13	209	68	293	210	53	317
14	613	661	1,793	647	373	2,260
15	272	162	1,220	274	191	1,168
16	394	15	384	408	49	278
17	200	61	245	201	48	265
18	498	237	1,233	500	195	1,311
19	106	8	26	107	6	29
20	421	299	617	424	221	683
21	624	69	4,015	885	601	3,097
22	389	10	3,660	844	652	2,446
23	582	221	947	582	207	966
24	442	61	2,680	626	520	2,073
25	165	156	270	186	72	388
26	296	163	660	299	121	730
27	227	25	181	227	25	182
28	1,402	308	1,002	1,403	271	1,043

TABLE 3 – ABSOLUTE INPUT PREDICTION ERRORS AND THE BEST INPUT MODEL DESIGNATIONS

Company	*CM-PM* (1)	*CPM* (2)	*CRV* (3)	Best Input Model (4)	Company	*CM-PM* (1)	*CPM* (2)	*CRV* (3)	Best Input Model (4)
1	.51	42	49	*CPM*	15	1 41	1.71	1.19	*CRV*
2	83	.42	39	*CRV*	16	30	1 84	38	*CM*
3	19	75	24	*CM*	17	34	66	.54	*CM*
4	54	03	1.14	*CPM*	18	1 28	.85	1 38	*CM*
5	81	44	94	*CPM*	19	.13	.52	.24	*CM*
6	1.05	8.35	.67	*CRV*	20	.21	.11	.34	*CPM*
7	1.58	27.01	1.10	*CRV*	21	67	16 93	.59	*CRV*
8	.25	1.07	49	*CM*	22	1 01	164 50	63	*CRV*
9	.31	38	.30	*CRV*	23	53	39	52	*CPM*
10	.35	.72	10	*CRV*	24	56	13 41	.30	*CRV*
11	1.29	.29	.69	*CPM*	25	08	.71	.11	*CM*
12	.12	.56	.31	*CM*	26	.90	.16	.74	*CPM*
13	1.28	.93	1.70	*CPM*	27	2.86	2 80	2.86	*CPM*
14	.60	.32	.53	*CPM*	28	.45	50	.41	*CRV*

TABLE 4 – ABSOLUTE OUTPUT PREDICTION ERRORS AND THE BEST OUTPUT MODEL, AND THE BEST OVERALL MODEL DESIGNATIONS

Company	*PM* (1)	*CPM* (2)	*CRV* (3)	Best Output Model (4)	Best Overall Model (5)
1	.782	.007	.007	*CPM-CRV*	*CPM*
2	.741	.147	.098	*CRV*	*CRV*
3	787	.010	.008	*CRV*	*CRV*
4	.780	247	.219	*CRV*	*CPM*
5	.839	294	292	*CRV*	*CPM*
6	731	.731	.624	*CRV*	*CRV*
7	.629	.276	.067	*CRV*	*CRV*
8	809	.016	003	*CRV*	*CRV*
9	769	382	.382	*CPM-CRV*	*CRV*
10	.760	.103	.055	*CRV*	*CRV*
11	691	033	077	*CPM*	*CPM*
12	.764	.016	.022	*CPM*	*CRV*
13	736	.014	.017	*CPM*	*CPM*
14	770	069	.018	*CRV*	*CPM*
15	618	463	460	*CRV*	*CRV*
16	.751	013	021	*CPM*	*CRV*
17	.739	.031	036	*CPM*	*CRV*
18	786	.170	.167	*CRV*	*CPM*
19	774	079	075	*CRV*	*CRV*
20	802	014	006	*CRV*	*CPM*
21	797	.409	.161	*CRV*	*CRV*
22	744	.509	066	*CRV*	*CRV*
23	.743	041	041	*CPM-CRV*	*CPM*
24	689	306	016	*CRV*	*CRV*
25	.670	078	042	*CRV*	*CRV*
26	.710	057	047	*CRV*	*CPM*
27	785	325	325	*CPM-CRV*	*CPM*
28	737	074	.074	*CPM-CRV*	*CRV*

where ϵ is assumed to be distributed $N(0,\sigma^2)$. Presumably, if the simulated solution is a good predictor of the actual, the parameter b should equal one. The regression estimate of b for the *CRV* model is .866. The null hypothesis that $b = 1$ cannot be rejected at the 20 percent significance level. In the case of the *CM-PM* model, the estimate is .799 and the null hypothesis is rejected at the 5 percent level. The hypothesis that $b = 1$ is rejected at the 1 percent level in the case of the *CPM* model.

It is worth noting that the extremely poor showing of the *CPM* model is a function of five "rate-base-maximizing" solutions which are singular to this model.[12] For example, if these five solutions are deleted, the average prediction error for the remaining firms is only .72. Interestingly, none of the firms in the sample implemented these rate-base-maximizing solutions even though the technologies are feasible.

C. *Comparing Output*

The predictive abilities of the models on the output side are evaluated in an identical fashion. The first three columns of Table 4 tabulate the absolute output prediction errors for each of the models while the fourth column designates the best output model. The *CRV* model predicts the best in eighteen cases with an average prediction error of .12. The *CPM* model predicts the best in six cases with an average prediction error of .18. In the remaining five cases, both the

[12] The rate-base-maximizing solutions require large K/H ratio technologies. These solutions have *CPM* prediction errors greater than 3 in Table 3.

CPM and *CRV* models predict equally well. The *PM* model is an inferior predictor with an average prediction error of .75.

Further proof of the superiority of the *CRV* model in predicting outputs is obtained by regressing for b' in the equation

$$Q_A = b'Q_S + \epsilon \tag{13}$$

where Q_A is the actual and Q_S the simulated output. Again, only the *CRV* model's estimate is close to 1, specifically, 1.041. The hypothesis that $b' = 1$ cannot be rejected at the 10 percent significance level. The identical hypotheses for the *PM* and *CPM* models are rejected at the 1 percent significance level

D. *Comparing Inputs and Output*

The constrained revenue-maximizing model is the best overall input-output predictor. The *CPM* model predicts output reasonably well, but not inputs, while the *PM* model predicts input proportions, but not output. The model which predicts inputs and output the best on a company by company basis can be determined by comparing both input and output prediction errors simultaneously. In fifteen cases the best overall model is unambiguous since the same model has the smallest input and output prediction errors. In the other thirteen cases the best model is resolved by comparing relative prediction errors assuming input and output prediction errors are weighted equally. For example, Colorado Interstate's *CPM* output prediction error is only .2 percent larger than the *CRV* output prediction error while the input prediction error of the former is 50 percent less than that of the latter. Consequently, the *CPM* model is the best overall predictive model for Colorado Interstate. Columne (5) in Table 4 lists the best overall predictive model in the sense just defined. The *CRV* model is seen to predict the best in seventeen cases, the *CPM* model in eleven cases and the *PM* model comes in a poor third.

E. *The Stability of the Results*

It should be noted that, with perhaps one exception, the simulated solutions and the conclusions in this section are quite robust with respect to a wide range of alternative parameter specifications.[13] The exception is the demand elasticity of sales for resale which is estimated to be 1.5. If demand is assumed to be less elastic than 1.5 the superiority of the *CRV* model over the *CPM* model is less pronounced, especially on the input side. This occurs because the number of rate-base-maximizing solutions in the *CPM* model appears to be a direct function of the size of the elasticity estimate. On the other hand, the *CRV* model's superiority is even more evident for elasticity specifications greater than 1.5.

III. Efficiency and Welfare Considerations

Showing that regulation has modified the behavior of the natural gas transmission industry addresses only one problem. To the policymaker, the critically important issue is whether or not regulation is beneficial to society at large. If rate of return regulation leads to input distortions, it also induces the firm to produce at a greater than profit-maximizing output level.[14] This tradeoff confronts the policymaker with the problem: do the benefits of regulation in the U.S. natural gas transmission industry outweigh the costs?[15] This issue can be resolved provided (i) the industry is valued by the social welfare function—the sum of producer's and consumer's surpluses—with its well-known deficiencies, and (ii) interpipeline rivalry is neglected.

The marginal cost pricing option (*MC*), which maximizes the social welfare function, is the benchmark against which the actual data and the simulated solutions are evaluated. The first column of Table 5 lists the maximum yearly benefits obtainable from a marginal cost pricing policy for each

[13] See the author, pp. 108-22. Included in the sensitivity analysis is the specification that ϕ is a constant elasticity supply function with a supply elasticity of 5.0.

[14] Except for the pathological case described by William Baumol and Alvin Klevorick.

[15] The policymaker could theoretically utilize this tradeoff to set socially optimal rates of return. See the author, G. Franklin Mathewson, and Herbert Mohring.

TABLE 5—GROSS SOCIAL WELFARE BENEFITS OF MARGINAL COST PRICING, THE 1965 ACTUAL, AND THE SIMULATED SOLUTIONS

Company	*MC* (1)	Actual (2)	*PM* (3)	*CPM* (4)	*CRV* (5)
1	123	122	92	122	122
2	159	154	118	149	152
3	221	220	166	220	220
4	182	173	124	166	168
5	112	112	82	110	110
6	413	400	309	308	335
7	999	919	744	859	915
8	95	95	72	95	95
9	247	243	184	229	229
10	78	77	59	75	77
11	333	317	250	316	321
12	161	160	123	159	160
13	111	106	81	108	108
14	436	420	312	413	421
15	512	454	367	401	401
16	403	396	303	394	396
17	114	111	86	111	112
18	371	359	268	356	357
19	33	32	25	32	32
20	270	264	188	263	264
21	829	792	567	711	777
22	745	712	542	595	722
23	345	332	252	335	335
24	610	565	440	519	565
25	150	140	110	135	141
26	253	241	187	239	240
27	471	467	359	453	453
28	453	443	336	446	447

Note Social welfare benefits are measured in millions of dollars per year

of the firms in my sample. The actual benefits and those obtainable from unconstrained profit, constrained profit, and constrained revenue maximization are presented in columns (2) to (5), respectively. These social benefits are gross of regulatory administrative costs. Net benefits can be estimated, albeit somewhat crudely, by subtracting $200 thousand of average yearly administrative costs per company from the gross figures.[16]

The public policy ramifications of Table 5 are straightforward. The effect of rate of return regulation on increasing output offsets the increased costs of input inefficiencies. Total actual net benefits for the industry are within 5.2 percent of the marginal cost pricing option, assuming the latter is effected costlessly. In all cases, net benefits are within 15 percent of the maximum. It is unlikely that other forms of politically acceptable regulatory procedures could do better.

APPENDIX A—AN ENGINEERING PRODUCTION FUNCTION

This appendix develops a Cobb-Douglas engineering production function for natural gas transmission. S.T. Robinson derives the following engineering production function for a compressor station and line pipe of length L miles:

$$(A1) \qquad Q = \frac{(.33)HPs^{.27}d^{1.8}}{L^{.36}}$$

where Q is output in cubic feet, HPs is station horsepower, d is the inside diameter of the line in inches, and the station has a dis-

[16] MacAvoy estimates that the *FPC* and the pipeline industry spent $3.5 million and $2.5 million, respectively, in 1968 or about $200 thousand per company.

charge pressure of 1000 *psi*. If the line is looped, L in (A1) is replaced by Le, the equivalent line length.[17] Assuming all loops are identical

$$Le = L/g^2 \tag{A2}$$

where L is now the length of one loop and g the number of loops. Therefore, (A1) becomes

$$Q = \frac{(.33)g^{.72}HPs^{.27}d^{1.8}}{L^{.36}} \tag{A3}$$

Assuming the pipeline is comprised of identically looped line sections, and noting that the number of compressor stations is (L^*/gL), (A3) can be manipulated to yield

$$Q = \frac{(.33)gHP^{.27}d^{1.8}}{L^{.09}L^{*.27}} \tag{A4}$$

where HP is total system horsepower and L^* is total system mileage.

The remainder of this appendix transforms (A4) into an engineering production function in the variables Q_1, H, and K. The variable K is substituted for d in (A4) using the equations:[18]

$$(D - d)/D = .0271 \tag{A5}$$

$$K = (7.05)(D^2 - d^2)L^* \tag{A6}$$

where D is the outside line diameter in inches.

Multiplying (A5) by $(D + d)$ gives

$$(D^2 - d^2) = (0.271)D(D + d) \tag{A7}$$

Substituting (A7) into (A6) and approximating D by d yields

$$K = (.382)d^2L^* \tag{A8}$$

Multiplying (A4) by $L^{*1.17}$, substituting (A8) into the result, and noting that $L^{-.09} = .7$, approximately,[19] yields

$$QL^{*1.17} = (.55)gHP^{.27}K^{.9} \tag{A9}$$

If the line operates at full capacity and g is treated as a parameter, (A9) becomes

$$Q = AH^{.27}K^{.9} \tag{A10}$$

where A denotes the scale constant.

The models are developed in terms of variable, not capacity output. Therefore, the constant in (A10) is (theoretically) adjusted by a capacity utilization rate and Q replaced by variable output which is approximated by:

$$(Q_1 + Q_2 + Q_3 - Q_4 + F + \Delta Qs) \tag{A11}$$

where F is compressor fuel consumption and ΔQs is the net change in natural gas storage inventories. If, in addition to the output variables, F and ΔQs are assumed to be proportional to Q_1, substituting (A11) into (A10) yields the Cobb-Douglas engineering technology function:

$$Q_1 = AH^{.27}K^{.9} \tag{A12}$$

Appendix B—The Data Base, and Some Variable and Parameter Estimates

The 1965 data base was chosen for three reasons. First, 1965 represents a long-run steady-state planning phase, the end of an extensive pipeline construction period characterized by new market penetration.[20] Pipeline construction from 1965 onwards is characterized by growth in existing markets and, therefore, a new planning phase. Second, peak load demand problems were absent in 1965. It is estimated that in that year 45 percent of all natural gas transmission lines operated at between 70 and 85 percent of capacity while the remainder were in the 85 percent plus range.[21] Third, disaggregated line-pipe capacity data are published for 1965.[22]

Most of the data are found in the *FPC* annual pipeline statistics, the *FPC* annual

[17] See American Gas Association, pp. 8/10–8/11.

[18] Equation (A5) assumes a discharge pressure of 1000 *psi* and a 5LX-52 high test line-pipe technology. See Donald Katz et al., pp. 628–30. Equation (A6) is derived by multiplying the volume of steel in an open cylinder by the weight of steel per unit of volume.

[19] Station spacing is usually (and optimally) constant for a given pipeline, although it differs from one pipeline to another. Typically L varies from 30 to 100 miles so that $L^{-.09}$ varies from .74 to .66.

[20] See *FPC Statistics* 1965, p. viii.

[21] See *FPC, National Gas Survey*, pp. 129–30.

[22] See the National Petroleum Council (*NPC*) report.

reports, *Moody's Public Utilities*, the *NPC* report on transportation capacities, and the John P. O'Donnell annual cost studies. These sources are sufficiently comprehensive to provide cross-sectional estimates for all but a few parameters. For the rest, industrywide estimates sufficed. Explanations are in order for some of the variable and parameter estimates.

The line-pipe capacity variable is measured in tons of main line steel. The *NPC* report provides a cross-sectional breakdown of pipeline mileage by outside diameters. Wall thicknesses are estimated by specifying an average industry steel technology, API Standard 5LX-52 with an operating pressure of 1000 *psi*.

Although peak load demand data are unavailable, the U.S. Bureau of Mines publishes a cross-sectional breakdown of main line industrial sales into interruptible and firm categories. Therefore, the proportion of peak load jurisdictional sales to total peak load sales can be estimated by $(Q_{1A}/Q_{1A} + \rho Q_{2A})$ where the A subscript denotes the 1965 actual outputs of Q_1 and Q_2 and ρ is the proportion of firm to total main line industrial sales.

The unit line-pipe capital cost is derived by summing the company's line-pipe related capital expenditures in the 1965 *FPC* accounts and dividing the result by K.[23] This book value figure is employed in the regulatory constraint and in the depreciation expense component of the objective function. In the remainder of the objective function, a 1965 constant dollar unit capital cost is used. This figure is obtained by adjusting all line-pipe related capital expenditures over the lifetime of the firm by a pipeline construction price index. The unit horsepower capacity capital cost is estimated in exactly the same fashion although the assumption of a constant unit cost is apparently contradicted by the potential economies of scale in horsepower generation. Average cost per horsepower generated declines with the capacity of a compressor-prime mover unit up to some technological limit In practice, there are severe limitations placed on the size of the compressor unit by the characteristics of the gas flow, the need for operational flexibility, and the dynamics of horsepower capacity utilization so that in 1965 average capacity was only from 1 to 2 thousand horsepower per unit.[24] Therefore, the assumption of a constant unit cost is not unreasonable. The wholesale price of the gas is estimated by average purchased gas expenditure. This figure is also the implicit price of pipeline produced gas.

The scale constants in the production function and the demand functions are estimated from the data and the appropriate functional forms. For example, the scale constant in the production function is determined by $(Q_{1A}/H_A^{.27}K_A^{.9})$ where the A subscript again denotes the 1965 actual values.

Accumulated depreciation in the regulatory constraint is expressed as a percentage of the undepreciated cost of the company's assets. This rate is calculated by dividing the *FPC* account "accumulated provisions for depreciation, amortization and depletion" by "total gas plant."[25]

A weighted average cost of capital is calculated for each company using the book value capitalization rates of debt, preferred and equity capital as of December 31, 1965. The pretax cost of debt is determined from Moody's rating of the most recent (pre-

[23] The assumption that line-pipe unit costs are constant is not unreasonable The following 1960-62 data and cost estimates give some indication of unit costs (on a per ton basis) for 24, 30, and 36" pipelines constrained to a working pressure of approximately 950 *psi*

Outside Diameter (")	Wall Thickness (")	Actual Data ($)	Norberg Estimates ($)	Columbia Gas Estimates ($)
24	.312	355	360	378
30	375	339	367	361
36	438		330	352

The Nordberg estimates are found in American Gas Association, p. 8/95. The actual cost data as well as the Columbia Gas estimates are from Laurence Rosenberg, p. 215.

[24] See the *NPC* report. There are modest economies of scale in the size of the compressor stations However, the cost of a reasonably sized station is proportional to the number and size of the compressor units and, therefore, horsepower capacity

[25] See *FPC Statistics*

1966) bonds issued by each pipeline company. The cost of preferred capital is taken to be the most recent (pre-1966) imbedded preferred share dividend yield. The after-tax cost of equity capital is derived from the familiar dividend yield equation where the growth rate is measured by the multiplicand of the retention rate and the return on (book value) equity capital, averaged over the period 1965–70. In those cases where the shares of the subsidiary pipeline company did not trade on the open market, the cost of capital of the parent is used.

The following parameter values are assumed to hold on an industry-wide basis. The fair rate of return is 6.5 percent which is the least upper bound on the return allowed pipeline companies under *FPC* jurisdiction from 1962 to 1967.[26] The straight line depreciation rates for K, H and other assets are 3.2, 3.9, and 4.5 percent, respectively. These depreciation rates are commonly employed in *FPC* rate case proceedings.[27] The long-run demand elasticity estimates are 1.5 for sales for resale and 4.0 for main line industrial sales. These elasticity estimates are borrowed from the MacAvoy and Noll study. It is assumed that 15 percent of interpipeline gas sales revenues are demand charges and the rest commodity charges. This figure is based on the original Atlantic Seaboard case[28] and appears to be the only estimate available. Fortunately, the simulations are insensitive to this particular parameter.

[26] See Stephen Breyer and MacAvoy, p. 31.
[27] See, for example, 13 *FPC* 53, 1954.
[28] See 11 *FPC* 521, 1952, and 11 *FPC* 57, 1952.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

W. J. Baumol and A. K. Klevorick, "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell J. Econ.*, Fall 1970, *1*, 162–90.

Stephen G. Breyer and Paul W. MacAvoy, *Energy Regulation by the Federal Power Commission*, Washington 1974.

J. L. Callen, "Production, Efficiency, and Welfare in the U.S. Natural Gas Transmission Industry," unpublished doctoral dissertation, Univ. Toronto 1976.

———, **G. F. Mathewson, and H. Mohring,** "The Benefits and Costs of Rate of Return Regulation," *Amer. Econ. Rev.*, June 1976, *66*, 290–97.

L. Courville, "Regulation and Efficiency in the Electric Utility Industry," *Bell J. Econ.*, Spring 1974, *5*, 53–74.

T. G. Cowing, "The Effectiveness of Rate-of-Return Regulation: An Empirical Test Using Profit Functions," in Melvyn Fuss and Daniel McFadden, eds., *Production Economics: A Dual Approach to Theory and Applications*, Amsterdam, forthcoming.

P. M. Hayashi and J. M. Trapani, "Rate of Return Regulation and the Regulated Firm's Choice of Capital-Labor Ratio: Further Empirical Evidence on the Averch-Johnson Model," *Southern Econ J.*, Jan. 1976, *42*, 384–98.

J. T. Jensen and T. R. Stauffer, "Implications of Natural Gas Consumption Patterns for the Implementation of End-Use Priority Programs," report to the Office of the General Counsel, General Motors Co., Arthur D. Little, Inc., 1972.

L. L. Johnson, "Behavior of the Firm Under Regulatory Constraint: A Reassessment," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 90–97.

Donald L. Katz et al., *Handbook of Natural Gas Engineering*, New York 1967.

P. W. MacAvoy, "The Effectiveness of the Federal Power Commission," *Bell J. Econ.*, Autumn 1970, *1*, 271–303.

——— **and R. Noll,** "Relative Prices on Regulated Transactions of the Natural Gas Pipelines," *Bell J. Econ.*, Spring 1973, *4*, 213–34.

J. P. O'Donnell, "Annual Study of Pipeline Installation and Equipment Costs," *Oil Gas J.*, various issues.

H. C. Peterson, "An Empirical Test of Regulatory Effects," *Bell J. Econ.*, Spring 1975, *6*, 111–26.

S. T. Robinson, "Powering of Natural Gas Pipelines," *J. Eng. Power*, A.S.M.E. Trans., July 1972, *94*, 181–86.

L. Rosenberg, "Natural Gas Pipeline Rate-Making Problems," unpublished doctoral dissertation, Cornell Univ. 1963.

R. M. Spann, "Rate of Return Regulation and Efficiency in Production: An Empirical Test of the Averch-Johnson Thesis," *Bell J. Econ.*, Spring 1974, *5*, 38–52.

S. H. Wellisz, "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *J. Polit. Econ.*, Feb. 1963, *71*, 30–43.

E. E. Zajac, "Lagrange Multiplier Values at Constrained Optima," *J. Econ. Theory*, Apr. 1972, *4*, 125–31.

American Gas Association, *Gas Engineers Handbook, New York 1965.*

Federal Power Commission (*FPC*), "Transmission Task Force Reports," *National Gas Survey*, 3 vols., Washington 1973.

———, *Reports*, various issues.

———, *Statistics of Interstate Natural Gas Pipeline Companies*, various issues.

Moody's Investors Service, *Moody's Public Utilities*, various issues.

National Petroleum Council (*NPC*), *U.S. Petroleum and Gas Transportation Capacities*, Washington 1967.

U.S. Bureau of Mines, *Main Line Natural Gas Sales to Industrial Users*, Washington 1974.

On the Optimal Provision of Journals qua Sometimes Shared Goods

By JANUSZ A. ORDOVER AND ROBERT D. WILLIG*

There are many interesting and important policy issues surrounding the provision of technical journals that arise from the simple fact that journals are sometimes shared goods:[1] that is, they are shared goods when offered to the reading public through libraries and are, at the same time, private goods to personal subscribers. It is said, for example, that publishers are experiencing increasing difficulty in recovering their "first copy costs" (setup costs) due to the rapid growth of reprography of library journals. Recognition of this new problem has led to intense public debate over copyright protection against uncompensated private dissemination of reproduced library materials.[2] There is a related accelerating trend towards the establishment of a dual pricing structure by publishers high rates for library subscriptions and lower rates for personal ones.

This paper shows that these policy issues can only be understood by viewing journals as sometimes shared goods. We provide a theoretical framework for the analysis of the optimal provision of such commodities. We then apply it to questions of journal subscription pricing and library usage fees.[3]

The theoretical framework is based on the simultaneous existence of separate but linked markets for private and shared use of the good. The private market serves those consumers who value the good above its price and who either have no access to a shared unit or who choose not to share the good because the cost saving to them is outweighed by the inconvenience. The users of shared units also fall into two groups. The first group is comprised of those who would not purchase a unit of the good for themselves even if the sharing option were not available to them. The second is the group of users of shared units of the good who would switch to private consumption if their sharing option were foreclosed. The sharing option enables the former group to enjoy services they otherwise would not, while it benefits the latter group by reducing their expenditures on the commodity.[4]

We establish that the critical linkage between markets for private and shared use of

*New York University and Bell Laboratories, and Bell Laboratories and Princeton University, respectively We would like to thank George Borts and the referee for helpful comments Ordovers's research was funded in part by a grant from the division of Science Information of the National Science Foundation

[1]We define a unit of a good as shared when its services are utilized by more than one agent, in contrast to a private unit of a good which is utilized by only one agent A commodity is a sometimes shared good, by our definition, if some units are shared and others are private. In our paper (1977b), we provide an economic analysis of the factors that determine whether a commodity will be utilized exclusively as a private good, exclusively as a shared good, or, lastly, as a sometimes shared good. A contrasting analysis of journals by Yoram Barzel rests on the public goods properties of the information disseminated in journals, while it ignores the public nature of library journal collections

[2]This debate was stimulated by the celebrated case of Williams and Wilkins Co vs the United States Summaries of various arguments and positions can be found in U.S. Senate *Hearings on S 1361*

[3]Questions pertaining to the choice of profit and welfare optimal subscription modes are discussed in Section IV of the authors (1977a)

[4]Users of shared units will, on average, have lower income than buyers of private units inasmuch as income is positively correlated with the inconvenience of shared use To this extent, it can be said that the purpose of sharing options is to benefit lower income groups In our welfare analysis below, we do not, however, place special weight on the benefits accruing to lower income consumers Neither do we consider any external benefits from the shared use of the commodity by those who could not afford private purchase (see the authors, 1976, for a discussion of optimal allocation when goods have both private and external benefits) Consequently, we are, perhaps, abstracting from major reasons for public support of options for shared use.

the good is this latter group of users. In our discussion of journals, they utilize the library copy when it is available to them because the price of personal subscription exceeds the inconvenience of patronizing the library. These potential subscribers, nevertheless, value the journal enough to buy a subscription were the library to discontinue its subscription.

We initially assume that libraries are perfect purveyors of the public good to their user populations (see Paul Samuelson). That is, they levy no use fee because there are no marginal costs of usage and they finance their acquisitions through lump sum fees which do not affect individuals' choices of reading modes. Further, a library subscribes to the journal if and only if the total willingness to pay of its population exceeds the institutional subscription price. In a marginal library, the willingness to pay just covers the library subscription price. The willingness to pay depends on the personal subscription price, and this further links the two markets. The financing and behavior of libraries is further discussed in Section I. There, too, we discuss the institutional assumptions and specify the formal model.

Section II studies the personal and institutional subscription prices that are optimal for profits and that are optimal for welfare under a binding nonnegative profit constraint. (The latter prices will be henceforth termed Ramsey-optimal prices.)[5] We assume that the production technology of journals exhibits increasing returns to scale to reflect the setup costs significant for public policy towards the publishing industry. Consequently, prices equal to marginal cost generate insufficient revenue to cover production cost, and thus violate the nonnegative profit constraint.

We see that the ratio of the optimal deviations of the two subscription prices from marginal cost depends on the ratio of the own-price elasticities of library and personal subscription demand, the ratio of library to personal subscriptions, and, the newly identified variable, the average number of potential personal subscribers who are users of the marginal libraries. However, the practical application of this result requires global information on the behavior of these critical functions of the prices. Unfortunately such data are unavailable.

Therefore, in Section III, we study the use of current values of the variables for the determination of the local price adjustments which are best for welfare, while leaving profit unchanged. The same expression for the ratio of the deviations of Ramsey-optimal prices from marginal cost can, when evaluated at current prices, be meaningfully compared with the ratio of current deviations. This expression is considerably simplified by the special structure imparted by the model to the market-demand elasticities. In particular, for reasonable and representative values of a small number of the current parameters, a journal currently setting equal personal and library subscription prices should move to a higher relative library price.

We apply these methods in a pilot study of the 1975 prices of five economics journals. We find that for four of them, welfare can be improved without loss of publisher profit by simultaneously increasing the library subscription price and decreasing the personal subscription price. Further, the hypothesis of profit maximization can be rejected for these journals.

In Section IV we study the economic impact of the introduction of library usage fees. It is theoretically surprising to discover that it is welfare suboptimal, under the nonnegative profit constraint, for the libraries to behave as perfect purveyors of their journal copies. While each population prefers to finance library subscriptions with lump sum taxes, they all benefit from collective adherence to a rule specifying that usage fees partially finance library acquisitions.

When the library usage fee is paid to the journal publisher, it can be interpreted as a copyright royalty. We show that under

[5] It was Frank Ramsey who first studied welfare-optimal prices under such a constraint. See William Baumol and David Bradford for a cogent survey.

weak and plausible conditions consumer welfare and profits can both be increased by the implementation of such a fee, when accompanied by appropriate decreases in the subscription prices. Thus, we identify the difficult policy problem of how to tie such price decreases to the extension of copyright protection to library usage. The Appendix contains the requisite proofs of our propositions.

I. The Model

We focus throughout on the provision of copies of a single journal by a monopolistic publisher.[6] The journal is assumed to be produced with an increasing returns-to-scale technology. This assumption captures the first copy costs and economies of scale in printing that are intrinsic to the publishing industry.[7] Moreover, it is this assumption that precludes the otherwise Ramsey-optimal marginal cost pricing of the good to both private and shared markets.[8]

We proceed with our analysis of Ramsey-optimal pricing under the working hypothesis that it is possible for the producer to set different prices in the two markets. In fact, journal publishers commonly follow this practice by setting distinct personal and institutional subscription prices. Such discrimination is facilitated by the repetitive and direct relationships between the publisher and the buyers. In markets for other sometimes shared goods where these factors are absent, the possibilities for price discrimination are more limited and the analysis would be more complex.[9]

In our modeling of library behavior, we continue to focus on a single journal and assume that the remainder of the library's collection is held fixed. We abstract from the operating and construction costs of libraries and from the concomitant overhead allocation problem. We posit that every group of potential readers has access to one already established and noncongested library facility.

We work with the simplest analytic model rich enough to represent the already introduced essential features of the interrelated private and institutional markets for journals as sometimes shared goods. Each individual agent is characterized by his gross benefits from consuming the good via the private and public modes, B and $B - T$, respectively. Thus T is the difference between the money-scaled subjective costs of utilizing the public and private modes. For example, T might measure the inconvenience of library use.

When an agent faces a personal subscription price of p_S, his net benefit from private subscription is $B - p_S$. When he faces a library usage fee of p_u, his net benefit is $B - T - p_u$ if he uses the library. The net benefit is 0 if he does not use the journal. Each agent selects the available mode that maximizes his net benefit. Thus, if he belongs to a group whose library does not own the journal, he will subscribe himself if and only if $B \geq p_S$. If, however, he does have access to a library copy, then he will buy a personal subscription if $B \geq p_S$ and $T + p_u > p_S$ $(B - p_S \geq B - T - p_u)$. He will be a library reader if $B - T - p_u \geq 0$ and $T + p_u < p_S$. Otherwise, he will choose not to read the journal.

It will be useful to dichotomize the library readers into the potential subscribers, for whom $B \geq p_S$ and $T + p_u < p_S$, and the

[6] We specifically ignore any possible monopolistically competitive interactions with publishers of other journals.

[7] This assumption is validated by the empirical findings in Baumol and Yale Braunstein, and in Braunstein (1976).

[8] See Baumol or John Panzar and Robert Willig for discussions of returns to scale and the feasibility of marginal cost pricing. It is shown in the authors (1977b) that marginal cost pricing, even when feasible, is not optimal for sometimes shared goods when the shared units are not perfectly purveyed.

[9] For example, price discrimination is limited by the incentives of the somewhat competitive intermediaries who sell such sometimes shared goods as books and consumer appliances. Yet it is sometimes possible for producers to sort out different classes of purchasers by offering similar products differentiated by quality. Thus book publishers can partially discriminate between library and private buyers by means of the sale of both hardcover and paperback editions.

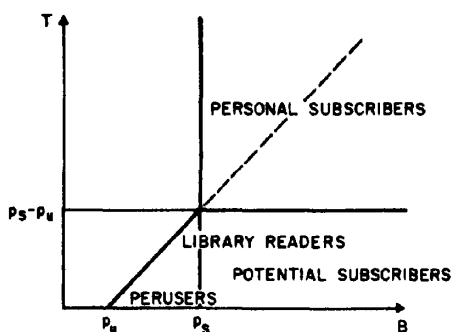



FIGURE 1

perusers, for whom $B < p_S$ and $T + p_u \leq B$. The latter group, unlike the former, would not buy personal subscriptions at p_S were the library to discontinue its subscription. Figure 1 depicts the aforementioned groups as regions in B, T space.

We assume that each consumer's indirect utility function is monotonically increasing in the sum of nominal income and maximized net benefits from journal use, holding fixed a numeraire price in another sector of the economy. Consequently, for each consumer, B and T are independent of any money expended for lump sum library taxes, use fees, or personal subscription payments.

We assume in Sections II and III that each library perfectly purveys its services to its population. In part, this requires that the library charge a usage fee equal to the marginal cost of usage. Henceforth, without loss of generality, this marginal cost is taken to be zero. Further, the financing of acquisitions is accomplished through a nondistorting mechanism.[10] Finally, acquisition decisions are made optimally for the aggregate welfare of the library population. Thus, the library will subscribe to a journal if and only if the aggregate willingness to pay, denoted by W, exceeds the institutional subscription price p_L.

We can express W as the difference between the population's aggregate net benefits with ($\overline{V}$) and without ($\underline{V}$) the library subscription, exclusive of the lump sum payments made to the journal publisher for the library subscription. Letting $h(B, T, m)$ denote the histogram function[11] of the population of agents served by the library with index m, we have

$$(1)\quad \overline{V}(p_S, m) = \int_{p_S}^{\infty}\int_{p_S-p_u}^{\infty}(B - p_S)h(B,T,m)\,dT\,dB + \int_{p_S}^{\infty}\int_{0}^{p_S-p_u}(B - T - p_u)h(B,T,m)\,dT\,dB + \int_{p_u}^{p_S}\int_{0}^{B-p_u}(B - T - p_u)\,h(B,T,m)\,dT\,dB$$

Reading left to right, the integrals measure the net benefits of the personal subscribers, the potential subscribers, and the persuers, respectively.

$$(2)\quad \underline{V}(p_S, m) = \int_{p_S}^{\infty}\int_{0}^{\infty}(B - p_S)h(B,T,m)\,dT\,dB$$

Here, without a library subscription, the only readers are personal subscribers. Finally,

$$(3)\quad W(p_S, m) = \overline{V}(p_S, m) - \underline{V}(p_S, m) = \int_{p_S}^{\infty}\int_{0}^{p_S-p_u}(p_S - T - p_u)h(B,T,m)\,dT\,dB + \int_{p_u}^{p_S}\int_{0}^{B-p_u}(B - T - p_u)\cdot h(B,T,m)\,dT\,dB$$

Thus, the personal subscribers contribute nothing to W, and the perusers are willing

[10] For example, the library subscription can be purchased with funds raised by a head tax levied on all people potentially served by the library, irrespective of whether or not a taxpayer actually uses the library facilities. Alternatively, we can assume that the requisite funds are raised with a perfectly discriminating tax on each library user which is less than his willingness to pay for the library services.

[11] Thus, $h(B,T,m)$ gives the density over B and T of the number of agents in the population of the library of type m. The function $h(\cdot)$ is finitely differentiable with respect to m.

to pay their full benefit, net of usage fee and inconvenience, $B - T - p_u$. However, the potential subscribers add only the difference $p_S - T - p_u$ between their evaluations of the library inconvenience together with the usage fee and the money cost of a personal subscription.

Using the willingness to pay concept, we can model the journal acquisition behavior of libraries. The marginal libraries, denoted by the index m^*, are just indifferent to acquiring the journal and thus satisfy

$$(4) \qquad W(m^*) = p_L$$

For convenience, we take m to be a scalar index defined so that the W function is increasing in m.[12] Then, letting the positive and differentiable function $f(m)$ denote the number of population groups with characteristic m,

$$(5) \qquad N^L = \int_{m^*}^{\infty} f(m)dm$$

is the total number of subscribing libraries.

The number of personal subscribers in a population m, with and without, respectively, a subscribing library is

$$(6) \quad \bar{N}^S(m) = \int_{p_S}^{\infty} \int_{p_S - p_u}^{\infty} h(B, T, m)dTdB$$

$$\underline{N}^S(m) = \int_{p_S}^{\infty} \int_{0}^{\infty} h(B, T, m)dTdB$$

Then, with p_u the same for all libraries, the total number of private subscribers is

$$(7) \quad N^S = \int_{0}^{m^*} \underline{N}^S(m) f(m)dm + \int_{m^*}^{\infty} \bar{N}^S(m) f(m)dm$$

In a subscribing library, the number of library readers is

$$(8) \; LR(m) = \int_{p_S}^{\infty} \int_{0}^{p_S - p_u} h(B, T, m)dTdB + \int_{p_u}^{p_S} \int_{0}^{B - p_u} h(B, T, m)dTdB$$

The total number of library readers is then

$$(9) \qquad LR^T = \int_{m^*}^{\infty} LR(m) f(m)dm$$

Publisher profit is the revenue from usage fees and subscriptions sales to individuals and institutions net of the cost of production

$$(10) \quad \Pi = p_u LR^T + p_L N^L + p_S N^S - C(N^S + N^L)$$

II. Ramsey-Optimal Subscription Prices

In this section, we characterize the personal and library subscription prices that maximize social welfare subject to nonnegative publisher profit. We proceed under the assumption that libraries are perfect purveyors of journals to their populations. Thus, with the postulated zero user costs, p_u is fixed at zero. Given our partial equilibrium framework, and abstracting from income distributional considerations, maximization of social welfare with respect to subscription prices is equivalent to the maximization of the sum of publisher profit (producer's surplus) and the aggregate net benefits to consumers from their journal use.[13] Using (1), (2), and (4), the aggregate net benefit to consumers is

$$(11) \; V = \int_{m^*}^{\infty} (\bar{V}(p_S, m) - p_L) f(m)dm + \int_{0}^{m^*} \underline{V}(p_S, m) f(m)dm$$

[12] In the authors (1977a) we show how to arrive at our results with a mathematically more satisfying representation of multidimensionally differentiated library populations.

[13] This equivalence follows from our earlier assumption that each individual's indirect utility function is monotonically increasing in the sum of income and net benefit from journal use. With such utility functions, there are no income effects on demands for journal use, and the net benefit is equal to the relevant consumer's surplus. This surplus is calculated as the line integral of the demands for both private subscriptions and for the use of the journal in the library.

Now, we can turn to the choice of the Ramsey-optimal p_S and p_L which maximize $V + \Pi$ subject to the constraint that $\Pi \geq 0$. Forming the Lagrangian, $L = V + \Pi + \lambda\Pi$, the analysis of the necessary first-order conditions for positive optimal prices requires that the partial derivatives of V with respect to p_S and p_L be calculated from the underlying model. Using (9), (1), and (2), we have

$$\frac{\partial V}{\partial p_L} = -\int_{m^*}^{\infty} f(m)dm - \frac{\partial m^*}{\partial p_L}[\overline{V}(m^*) - \underline{V}(m^*) - p_L]\, f(m^*)$$

However, because of the definitions of m^*, W, and N^L, (3), (4), and (5), the second term is zero and we are left with $\partial V/\partial p_L = -N^L$. Similar calculations yield $\partial V/\partial p_S = -N^S$.[14]

These results, together with routine differentiation of the profit function (10) with respect to p_S and p_L, allow us to conclude from the necessary optimality conditions:

PROPOSITION 1: *The Ramsey-optimal subscription prices satisfy*

$$(12)\quad \begin{pmatrix} p_L - c \\ p_S - c \end{pmatrix} = \frac{-\lambda}{\lambda + 1}\left(\frac{1}{N^L_L N^S_S - N^S_L N^L_S}\right) \cdot \begin{pmatrix} N^S_S & -N^S_L \\ -N^L_S & N^L_L \end{pmatrix}\begin{pmatrix} N^L \\ N^S \end{pmatrix}$$

or

$$(13)\quad \rho \equiv \frac{p_L - c}{p_S - c} = \frac{N^S_S N^L - N^S_L N^S}{-N^L_S N^L + N^L_L N^S} \equiv \psi$$

where c denotes the marginal cost of journal publication $C'(N^S + N^L)$, and subscripts S and L denote partial derivatives with respect to p_S and p_L.

Equations (12) are the standard Ramsey rule for optimal deviations of prices from marginal costs under the nonnegative profit constraint.[15] In the present form, (12) is not very illuminating. A more useful formulation can be derived by substituting into it detailed relationships among the partial derivatives of demand extracted from the underlying model.

PROPOSITION 2: *Let Z be the number of potential subscribers in each marginal library. Then $N^L_l < 0$ and*

$$(14)\quad \infty > N^S_L = N^L_S = -ZN^L_L \geq 0$$

Further, let $\hat{N}^S_S$ denote the derivative of N^S with respect to p_S holding constant the set of subscribing libraries. Then

$$(15)\quad -\infty < N^S_S = \hat{N}^S_S - N^L_S Z \leq \hat{N}^S_S < 0$$

Equation (15) indicates that the own-price sensitivity of personal subscriptions can be decomposed into two effects. The first is the regular substitution effect with the extent of the institutional facilities available to consumers held fixed. The second reflects the reaction in individuals' private demands induced by the endogenous changes in the library facilities available to them.

Equation (14) says that library and private journal subscriptions are weak gross substitutes. Demand for personal subscriptions is affected by p_L through the sensitivity to this price of the set of subscribing libraries. When a marginal library drops its subscription, its potential subscribers enter the market for private subscriptions. The reverse flow occurs if a decrease in p_L induces a marginal library to acquire a journal copy.

Demand for library subscriptions is affected by p_S through the sensitivity to this price of the willingnesses to pay. For example, an increase in p_S reduces the net benefit of personal subscription and hence increases the value of the library copy to the

[14]These results are analogous to Roy's Law for an individual utility-maximizing consumer (see Donald Katzner for a clear exposition). Thus, in our model aggregate consumer and library demands behave as if they maximize an aggregate utility function subject to an aggregate budget constraint. This result is surprising because the demand for library subscriptions is determined by the simultaneous collective decisions of many population groups, while N^S results from the individual decisions of the agents. The result hinges critically on the welfare optimal acquisition policy of libraries.

[15]See Marcel Boiteux, for example.

potential subscribers; that is, to those who use the library copy and to whom private subscription is a viable alternative. Thus, for a change in the price of either personal or library subscriptions, the critical linkages between the two markets are formed by the set of potential subscribers who use each marginal library. Equation (14) states that in our model the linkages are in fact symmetric.

If there were no potential subscribers, the markets for private and library subscriptions would be decoupled, the cross elasticities of demand would be zero, and the optimal prices would be characterized by the inverse elasticity rule. Otherwise, the role played by the number of potential subscribers in the determination of the optimal prices is exposed by substituting (14) into (13).

PROPOSITION 3: *The Ramsey-optimal prices are characterized by*

(16)

$$\frac{p_L - c}{p_S - c} \equiv \rho = \psi \equiv \frac{[N_S^S/N^S]/[N_L^L/N^L] + Z}{1 + (N^L/N^S)}$$

It is evident from (16) that $\psi > 0$, implying that at the optimum $p_L - c$ and $p_S - c$ have the same sign. The assumed increasing returns to scale rule out the possibility that $p_S \leq c$ and $p_L \leq c$, since such prices would violate the nonnegative profit constraint. Thus, Ramsey-optimal subscription prices must both exceed marginal cost.

In Section III we shall employ (16) to study the desirability of the relationships between current journal prices. However, here, we can combine equations (15), (14), and (13) to gain insight into the relationship between the optimal levels of p_L and p_S.

PROPOSITION 4: *The ratio of the Ramsey-optimal deviations between prices and marginal cost are bounded as follows:*

$$Z < \rho \leq Z + \frac{\tilde{N}_S^S N^L}{N_L^L N^S} \tag{17}$$

Thus, $Z \geq 1$ would immediately imply that $\rho > 1$, that the optimal library price exceeds the optimal personal subscription price.[16]

III. Determining Best Price Adjustments from Current Data

There is considerable methodological difficulty in deriving insights from (16) that are relevant to current practices of journal pricing. The variables (elasticities, circulations, and number of potential subscribers) to which the formula relates ρ are all to be evaluated at yet to be determined prices. This endogeneity, endemic to necessary first-order conditions, means that the optimal prices can only be determined as the solutions to simultaneous equations whose global behavior is almost impossible to deduce from available local data. Further, intuitions that we may have concerning *current* values of the variables governing ρ cannot be logically utilized via such first-order conditions as (16) to illumine the optimal prices.

Fortunately, there is an analytic line of inquiry which circumvents these conceptual difficulties. We can ask for the *direction* of change from the current prices which is best for social welfare while preserving the current level of profit. It can be shown[17] that if the current $\rho = (p_L - c)/(p_S - c)$ is greater than the current value of ψ (defined in (13)), then the best profit-constrained direction of changes requires that p_L be lowered and p_S be raised. Inversely, if, at current levels $\rho < \psi$, then p_L should be raised and p_S lowered. It should be emphasized that these calculations do not necessarily indicate the relationships between the current and the optimal prices. Instead, they give the best local price adjustments that can be determined from

[16]It should be noted that when library populations are described by more than one characteristic, all marginal libraries are not identical and Z is properly interpreted as the average number of potential subscribers in the marginal libraries. In such more realistic models, noninteger values of Z have meaningful interpretations. See fn. 12.

[17]See Willig and Elizabeth Bailey

strictly local information on the relevant functions.

From this point of view, ψ, calculated at current values of the variables, can indeed be meaningfully compared with the current ratio $(p_L - c)/(p_S - c)$. Since (16) gives an expression equal to ψ, it can serve as a vehicle for the application of current data to the study of present journal prices, yielding recommendations for the best direction of change. Further, we can study the level of ψ, always evaluated at current prices, as a function of the values its parameters could take on as they pertain to different journals.

We shall first utilize this technique to establish conditions under which it can be asserted that welfare would increase (without affecting profits) by introducing a positive margin between currently equal library and personal subscription prices. This assertion can be made if the current value of ψ for a particular journal with $p_S = p_L$ exceeds 1. For this journal, the current ρ is equal to 1, less than ψ, indicating that p_L should be raised and p_S lowered.

For notational convenience, let $n = N^L/N^S$ and let k be the ratio of the own-price elasticity of demand for private subscriptions to the own-price elasticity of demand for library subscriptions. In this notation, (16) becomes:

$$\psi = \left(\frac{p_L}{p_S} k + Z\right)/(1 + Zn) \tag{18}$$

PROPOSITION 5: *If, currently, $p_S = p_L$, then raising p_L and lowering p_S would increase social welfare and leave publisher profit unchanged whenever $(k - 1) + Z(1 - n) > 0$.*

This condition will be met whenever the circulation ratio n is less than 1, and the ratio of the elasticities k is greater than 1. The meager empirical evidence suggests that k is significantly larger than 2, for all journals studied.[18] Further, the best available data indicates that $n < 1$ for a majority of technical journals.[19] Thus a finding that $\psi > 1$ for a journal with $p_S = p_L$ would not be surprising, and the policy recommendation to differentiate the subscription prices, $p_L > p_S$, would be forthcoming from our model.

For journals already charging differentiated prices, the investigation of the best direction of price changes requires more current information. If k, n, and Z were known then the test is just $\rho \gtrless \psi$. However, Z may be more difficult to estimate than are k or n. Nevertheless, we can use (18) to determine the minimum value of ψ, over all $Z \geq 0$, as a function of k and n. If for a particular journal it should be the case that $\rho < \psi_{min}$, then surely $\rho < \psi$ and the recommendation to increase p_L and decrease p_S would follow.

PROPOSITION 6:

$$\psi \geq \psi_{min} \equiv \begin{cases} (p_L/p_S)k & \text{if } (p_L/p_S)nk \leq 1 \quad \text{(19a)} \\ \dfrac{\dfrac{p_L}{p_S} k + \left[\dfrac{kp_L}{np_S}\right]^{1/2}}{1 + n\left[\dfrac{kp_L}{np_S}\right]^{1/2}} & \text{if } (p_L/p_S)nk > 1 \quad \text{(19b)} \end{cases}$$

and ψ_{min} is an increasing function of k.

Because we do not have a precise estimate of k, the fact that ψ_{min} is monotonically increasing in k facilitates the analysis of current journal subscription prices. First, it can be inferred that $\rho < \psi$ if ψ_{min} calculated at an underestimate of k exceeds ρ. For this purpose, we continue to assume that $k \geq 2.0$. Second, we can find that value of k, denoted by k^*, at which $\rho = \psi_{min}$.

PROPOSITION 7:

$$k^* = \begin{cases} \dfrac{p_S}{p_L}\rho & \text{if } (p_L/p_S)nk \leq 1 \quad \text{(20a)} \\ \dfrac{p_S}{p_L} n\rho^2 & \text{if } (p_L/p_S)nk > 1 \quad \text{(20b)} \end{cases}$$

[18]See Sanford Berg. Braunstein (1977) seems to indicate values of k significantly above 2

[19]See Bernard Fry and Henry White

TABLE 1

1975	p_S	p_L	$N^L + N^S$	N^L/N^S	Pages	c	$\frac{p_L - c}{p_S - c}$	$\psi_{min}(k = 2)$	k^{*b}
QJE	15.00	15.00	5,500	1.22	686	7.85	1.00	1.28	1.22
AER[a]	23.00	34.50	27,500	.37	3,280	15.18	2.47	2.85	1.50
JPE	15.00	20.00	8,400	.43	1,318	11.51	2.43	2.49	1.90
EI	14.00	20.00	3,800	.49	610	8.32	2.06	2.42	1.46
JET	34.50	69.00	1,500	4.30	873	17.05	2.98	.97	19.09

[a]Includes *Papers and Proceedings* and the *Journal of Economic Literature*
[b]The value of k for which $\psi_{min} = \rho$.

If the true value of k were greater than k^* then Proposition 5 would imply that $\rho < \psi$. Knowledge of k^* enables us to ascertain how small a value of k would support the policy conclusion to raise p_L and lower p_S.

We now apply these methods in a pilot study of the 1975 prices of five economics journals: *Quarterly Journal of Economics* (*QJE*); *American Economic Review*, together with the *Papers and Proceedings* and the *Journal of Economic Literature* (*AER*), *Journal of Political Economy* (*JPE*); *Economic Inquiry* (*EI*); and the *Journal of Economic Theory* (*JET*). The prices, taken from the public record, pertain to all issues published in 1975. For the association journals (*AER* and *EI*), we took p_S to be the membership fee, and we ignore any benefits and costs of membership unrelated to the journal subscriptions. Circulation figures N^L and N^S were obtained directly from the editorial offices.[20] The marginal costs were calculated from the formula,[21]

$$\ln \frac{C(Q)}{10{,}000} = -.742 + .559 \ln \frac{N^S + N^L}{1000} + .814 \ln \frac{\text{annual editorial pages}}{100}$$

and then inflated by 25 percent.[22] These data appear in columns 1-6 of Table 1. Column 7 holds ρ, the ratio of the deviations of the subscription prices from marginal cost, which is to be compared with ψ.

For each of the five journals, $(p_L/p_S)nk > 1$ for $k \geq 2$, and so we can presume that (19b) applies. Column 8 lists the values of ψ_{min} computed from (19b) with the underestimate of 2.0 used for k. Column 9 exhibits k^* calculated from (20b), the value of k which would make $\psi_{min} = \rho$.

These calculations suggest that ρ is indeed well below ψ for all the journals but *JET*. Both intuition and the evidence support the contention that the own-price elasticity of personal subscriptions is more than twice that of library subscriptions. With $k > 2$, both columns 7 and 8 show that the values of ρ are below those of ψ_{min}. The policy conclusion[23] is that net consumer welfare can be increased, while the levels of publishers' profits are maintained, by simultaneously increasing p_L and decreasing p_S, for *QJE*, *AER*, *JPE*, and *EI*.

For *JET*, Table 1 shows that it is unlikely that $\rho < \psi_{min}$. Since $nk(p_L)/(p_S) > 1$, ψ is decreasing in Z, and (18) yields $\psi_{max} =$

[20]These data for *AER* and *EI* are annually released publicly. The editorial offices of *JPE* and *JET* specified $N^S + N^L$ precisely, and offered estimates of N^L/N^S. The editorial office of *QJE* offered estimates of 1975 $N^S + N^L$ and N^L/N^S

[21]This equation is presented and described in Baumol and Braunstein.

[22]The equation appears to be based on 1973 factor prices. We assume here that the prices of all costly factors of journal production rose by 25 percent between 1973 and 1975. Both the Wholesale Price Index of book paper and the Bureau of Labor Statistics index of printing trades wages did increase by approximately 25 percent between those dates.

[23]Of course, the conclusions rest upon the empirically untested model, and upon the numbers presented in Table 1. We regard this as a pilot study, hopefully pointing the way towards a full empirical treatment of both the model and the relevant parameters. Note that (14) and (15) can be utilized to generate several testable implications of the model.

$k(p_L)/(p_S)$. Since $(p_L/p_S) = 2$, $\rho < \psi_{max}$ for $k > 1.5$. Thus, for reasonable values of k, $\psi_{min} < \rho < \psi_{max}$, and we cannot reject the hypothesis that the subscription prices of *JET* satisfy the optimality conditions. In fact, rearrangement of (18) shows that $\rho = \psi$ if k and Z satisfy $k = 1.5 + 6Z$. It is certainly plausible, for example, that $Z = .5$ and $k = 4.5$.

Thus far we have studied welfare maximization, and our concern with profits has been restricted to the constraint of nonsubsidized viability of the publisher. However, these very same concepts can also be usefully applied to the study of profit maximization.

PROPOSITION 8: *A necessary condition for the current levels of p_L and p_S to be profit maximizing is that $\rho = \psi$.*

Thus the results displayed in Table 1 can be interpreted as evidence that all the journals but *JET* are neither successful profit maximizers nor constrained welfare optimizers.

IV. Library Usage Fees

In this section we study the welfare effects of the introduction of library usage fees. First we consider the imposition of user charges as a partial substitute for the lump sum taxes previously assumed to completely finance library acquisitions. Second, we analyze user fees which are paid directly to the publisher. These payments can be interpreted as copyright royalties.

Consider a rule that each library must finance the proportion α of the subscription price p_L by means of a use fee p_u. For the library of type m, this rule implies that

$$p_u(m)LR(m) = \alpha p_L \tag{21}$$

Here, $LR(m)$ given by (8) is the number of readers using library m when the usage fee is $p_u(m)$. In view of (21) and the heterogeneity of libraries, the usage fee varies among libraries.

Under this regime, with $\alpha > 0$, libraries do not perfectly purvey journals since the positive user fee exceeds the zero marginal cost of usage. Each library population as a group would prefer to circumvent the positive α rule and to pay p_L solely out of the lump sum taxes characteristic of perfect purveyance. However, in the Appendix, we prove the following surprising result:

PROPOSITION 9: *The introduction of a positive α increases aggregate consumer welfare with Ramsey-optimal subscription prices* (12) *above marginal cost, whenever there are any marginal prospective subscribers in any of the subscribing libraries.*

The intuitive explanation is that as α and consequently the user fees are infinitesimally raised from zero, the potential subscribers who were indifferent between library use and personal subscription are induced with no welfare loss to subscribe. This raises publisher profit by approximately $p_S - c$ for each new private subscription. Because the profit constraint is binding, the profit increment enables p_S and p_L to be lowered, bringing profit back to its original level, and increasing consumer welfare. The import of Proposition 9 is that perfect purveyance of shared units of a sometimes shared good is not generally optimal. For small usage fees employed to partially finance library acquisitions, the positive effect on publisher profit from increased private subscriptions outweighs the undesirable effects on the allocation of journals to libraries and on individuals' choices of reading modes.

We now show that a more practical system of positive usage fees which are uniform across libraries and paid directly to publishers is also generally desirable. The propositions that follow are proven in the Appendix from the model specified in Section I.

PROPOSITION 10: *Suppose that at the profit-maximizing p_S and p_L, with $p_u = 0$,* (i) *there are some marginal potential subscribers in some subscribing libraries and* (ii) *the number of journal readers in each marginal library is less than the average number of journal readers in all the subscribing*

libraries; i.e.,

$$LR^T/N^L > LR(m^*) \tag{22}$$

Then, there exist p_u, p_S, p_L, with $p_u > 0$, at which both publisher profit and consumers' welfare are greater than they are with $p_u = 0$ and with p_S, p_L set profit maximally.

The rationale of this result is that the introduction of a positive usage fee paid to the publisher directly increases his revenues, and in addition further increases his profit through the induced rise in demand for personal subscriptions. The proposition asserts that subscription prices can be lowered sufficiently to compensate consumers collectively for the positive user charge without completely nullifying the gain in profit.

Nevertheless, it is problematic whether the profit-maximizing publisher would find it in his own interest to effect these requisite price reductions if he were to be granted the right to collect a usage fee. In response to the increase in demand for personal subscriptions resulting from the newly positive p_u, the publisher may well find it profit optimal to raise p_S. In short, consumer welfare may be lowered by allowing a profit-minded publisher to charge a usage fee, even if its level is set by the government.

In contrast, consumer welfare is improved by the introduction of a usage fee when p_S and p_L are chosen optimally for net welfare subject to the nonnegative profit constraint.

PROPOSITION 11: *Suppose that at the Ramsey-optimal p_S and p_L, with $p_u = 0$, (22) holds. Then, there exist p_S, p_L, p_u with $p_u > 0$, at which publisher profit is unchanged while consumers' welfare is greater than it is in the situation above.*

Overall, we find that a copyright based library usage fee is a practically feasible instrument[24] which is desirable when properly employed. For profit-maximizing publishers, such extension of copyright protection can increase both profit and consumers' welfare. The challenge for public policy is to develop an institution tying price reductions that benefit consumers to copyright protection that increases publishers' profits. The present analysis shows the existence of such a compromise pricing package that will benefit both publishers and readers.

Further, we have shown that a usage fee is a beneficial instrument in the hands of a welfare minded price setter. It can be conceivably argued that nonprofit journal publishers do, in fact, seek (albeit unsuccessfully, as indicated in Section III) to set prices in this way. Then, for this major category of publishers, our results may be interpreted to recommend library usage charges.

APPENDIX

PROOF of Proposition 1:

The first-order conditions for the maximization with respect to p_S and p_l of $L = V + \Pi + \lambda\Pi$ are

$$\frac{\partial L}{\partial p_S} = -N^S + (\lambda + 1) \cdot [N^S + (p_S - c)N^S_S + (p_L - c)N^L_S] = 0$$

$$\frac{\partial L}{\partial p_l} = -N^l + (\lambda + 1) \cdot [N^L + (p_S - c)N^S_L + (p_L - c)N^L_L] = 0$$

Solving these equations for $p_S - c$ and $p_L - c$ by matrix inversion yields (12).

PROOF of Proposition 2:

Working from (5), we obtain

$$N^l_L = -\frac{\partial m^*}{\partial p_L} f(m^*) \tag{A1}$$

[24]This payment mechanism can be implemented if the publisher can monitor total usage of library copies. See U.S. Senate *Hearings on S.1361* for arguments concerning the feasibility of such monitoring. Note that when the library population must collectively pay p_u per each journal use, it has the incentive to in fact levy a charge of p_u on each user. This should be contrasted with the incentive structure of the mechanism defined in (21). Each library population has the incentive to circumvent that scheme by waiving the required user fees and by financing p_L completely through lump sum taxes. The publisher cannot ascertain by monitoring usage whether such circumvention has occurred.

and $N^L_S = -\frac{\partial m^*}{\partial p_S} f(m^*)$

Implicit differentiation of (4) gives

(A2)

$$\frac{\partial m^*}{\partial p_L} = 1\Big/\frac{\partial W}{\partial m} \text{ and } \frac{\partial m^*}{\partial p_S} = -\frac{\partial W/\partial p_S}{\partial W/\partial m}$$

Thus, $N^L_L = -f(m^*)/(\partial W/\partial m)$ which is negative by construction. Differentiation of (3) yields

(A3)

$$\frac{\partial W}{\partial p_S} = \int_{p_S}^{\infty} \int_0^{p_S - p_u} h(B,T,m^*)dTdB \equiv Z$$

This is the number of potential subscribers who frequent each marginal library. Together, (A1), (A2), and (A3) yield $N^L_S/N^L_L = -\partial W/\partial p_S = -Z$.

Differentiating (7) with respect to p_L, we see

$$\text{(A4)}\; N^S_L = \frac{\partial m^*}{\partial p_L} f(m^*)[\underline{N}^S(m^*) - \bar{N}^S(m^*)]$$

Using (6) and the definition of Z in (A3), (A4) becomes

$$N^S_L = \left(\frac{\partial m^*}{\partial p_l}\right) f(m^*)Z$$

Then, using (A1), $N^S_L = -N^L_L Z = N^L_S$.

Differentiating (7) with respect to p_S gives

$$\text{(A5)}\quad N^S_S = \frac{\partial m^*}{\partial p_S} f(m^*)(\underline{N}^S(m^*) - \bar{N}^S(m^*)) + \left[\int_0^{m^*} \underline{N}^S_S(m)f(m)dm + \int_{m^*}^{\infty} \bar{N}^S_S(m)f(m)dm\right]$$

We denote by $\hat{N}^S_S$ the negative terms in the brackets which represent the derivative of N^S with respect to p_S, holding constant the set of subscribing libraries. Using (A2), (A3), and (A4), we have

$$\frac{\partial m^*}{\partial p_S} f(m^*)(\underline{N}^S(m^*) - \bar{N}^S(m^*)) = -N^L_S Z$$

Thus, (A5) can be rewritten

$$N^S_S = -N^L_S Z + \hat{N}^S_S \leq \hat{N}^S_S < 0$$

The demand derivatives are finite because all component functions are assumed to be finitely differentiable.

PROOF of Proposition 4:

At the Ramsey optimum, by (13) and (15),

$$\rho = \frac{N^S_S N^L - N^S_L N^S}{-N^L_S N^L + N^L_L N^S} = \frac{N^L[\hat{N}^S_S - N^L_S Z] - N^L_S N^S}{-N^L_S N^L + N^L_L N^S}$$

Then, using (14),

$$\rho = Z + \frac{\hat{N}^S_S}{N^L_L}\left[\frac{N^L/N^S}{1 + ZN^L/N^S}\right]$$

Since

$$0 < \frac{\hat{N}^S_S}{N^L_L}\left[\frac{N^L/N^S}{1 + ZN^L/N^S}\right] \leq \frac{\hat{N}^S_S}{N^L_L}\frac{N^L}{N^S}$$

(17) follows.

PROOF of Proposition 5:

With $p_S = p_L$, (18) shows that $\psi = (k + Z)/(1 + Zn)$, and $\rho = 1$. Hence $\psi > \rho$ if $k + Z > 1 + Zn$, or if $(k - 1) + Z(1 - n) > 0$.

PROOF of Proposition 6:

Differentiation of (18) shows that the expression for ψ is either monotone increasing or decreasing in Z as $(p_L/p_S)nk$ is less or greater than 1. In the former case, setting Z at its lower bound of 0 gives $\psi_{min} = (p_L/p_S)k$. In the latter case, we require an upper bound on Z to calculate ψ_{min}.

Together, (14) and (15) yield

$$0 > \hat{N}^S_S = N^S_S + N^L_S Z = N^S_S - Z^2 N^L_L$$

Thus $$Z^2 < N^S_S/N^L_L = \frac{k}{n}\frac{p_L}{p_S}$$

and $$Z < \sqrt{\frac{k}{n}\frac{p_L}{p_S}}$$

Substituting this upper bound for Z into (18) gives (19b).

Differentiation of (19a) and (19b) shows that they are increasing functions of k.

PROOF of Proposition 7:

Equating ρ to (19a) trivially yields (20a). Equating ρ to (19b) and rearranging gives

$$\frac{p_L}{p_S}k^* - \rho = (\rho n - 1)\left[\frac{k^* p_L}{n p_S}\right]^{1/2}$$

Squaring both sides, rearranging, and factoring reveals

$$\left(k^* - \frac{n\rho^2 p_S}{p_L}\right)\left(\frac{p_L}{p_S}nk^* - 1\right)\frac{p_L}{p_S n} = 0$$

Since, in this case, $(p_L/p_S)nk > 1$, (20b) follows.

PROOF of Proposition 8:

The first-order conditions for profit maximal p_S and p_l are:

$$\frac{\partial \Pi}{\partial p_S} = N^S + (p_S - c)N^S_S + (p_l - c)N^L_S = 0$$

$$\frac{\partial \Pi}{\partial p_l} = N^L + (p_S - c)N^S_L + (p_l - c)N^L_L = 0$$

Solving these equations for $p_S - c$ and $p_L - c$ yields (13), (14), and $\rho = \psi$

PROOF of Proposition 9:

Note first that given (21), it is $(1 - \alpha)p_L$ that must be covered by a library's willingness to pay. Thus,

$$\text{(A6)} \qquad W(m^*) = (1 - \alpha)p_L$$

Differentiation with respect to α yields

$$\text{(A7)} \quad \frac{dm^*}{d\alpha} = \frac{-\dfrac{\partial W(m^*)}{\partial p_u}\dfrac{\partial p_u(m^*)}{\partial \alpha} - p_L}{\dfrac{\partial W(m^*)}{\partial m} + \dfrac{\partial W(m^*)}{\partial p_u}\dfrac{\partial p_u(m^*)}{\partial m}}$$

Differentiation of (3) and use of (8) shows that

$$\text{(A8)} \qquad \frac{\partial W(m)}{\partial p_u} = -LR(m)$$

Differentiation of (21) shows that, at $\alpha = 0$, $\partial p_u(m)/\partial \alpha = p_L/LR(m)$. Substituting this and (A8) into (A7) establishes that at $\alpha = 0$, $\partial m^*/\partial \alpha = 0$.

Differentiation of (11) with respect to α, recognizing that now p_u is the function of both α and m given by (21), shows that at $\alpha = 0, \partial V/\partial \alpha = 0$.

Profit is now simply $\Pi = p_S N^S + p_L N^L - C(N^S + N^L)$, where N^S and N^L are given by (5) and (7), again remembering that (21) and (A6) give new interpretations to p_u and m^*. Here, in view of the fact that $\partial m^*/\partial \alpha = 0$, calculation shows that at $\alpha = 0$,

$$\text{(A9)} \quad \frac{\partial \Pi}{\partial \alpha} = (p_S - c)\cdot \int_{m^*}^{\infty}\frac{\partial p_u(m)}{\partial \alpha}\int_{p_S}^{\infty} h(B, p_S, m)f(m)dBdm = (p_S - c)p_L\int_{m^*}^{\infty}\frac{1}{LR(m)}\cdot \int_{p_S}^{\infty} h(B, p_S, m)f(m)dBdm$$

The number of marginal potential subscribers in library m when $p_u = 0$ is given by

$$\int_{p_S}^{\infty} h(B, p_S, m)dB$$

Thus, under the hypotheses of the proposition, at $\alpha = 0, \partial \Pi/\partial \alpha > 0$.

Let $V^* + \Pi^*$ denote the Ramsey-optimal level of net social welfare. Viewing α as a parameter of the Ramsey-optimization program, the envelope theorem implies that $d(V^* + \Pi^*)/d\alpha = \partial V/\partial \alpha + (1 + \lambda)(\partial \Pi/\partial \alpha)$, where λ is the positive Lagrange multiplier on the profit constraint and where the derivatives are evaluated at the optimal p_S and p_L. Since $\partial V/\partial \alpha = 0$, and since $d\Pi^*/d\alpha = 0$ because of the profit constraint, $dV^*/d\alpha = (1 + \lambda)(\partial \Pi/\partial \alpha) > 0$, at $\alpha = 0$. This proves the proposition.

PROOF of Proposition 10:

Consider an infinitesmal increase in p_u from 0 accompanied by the decrease in p_L which keeps m^* unchanged. Along this path,

$$\text{(A10)} \qquad \left.\frac{dp_L}{dp_u}\right|_{m^*} = -\frac{\partial m^*/\partial p_u}{\partial m^*/\partial p_L}$$

Calculating from (4) and (A8),

$$\text{(A11)} \qquad \frac{\partial m^*}{\partial p_u} = \frac{LR(m^*)}{\partial W(m^*)/\partial m}$$

Using (A2) and (A11), (A10) becomes

$$\text{(A12)} \qquad \left.\frac{dp_l}{dp_u}\right|_{m^*} = -LR(m^*)$$

Differentiating (10) along this path by the chain rule yields

$$\text{(A13)} \quad \left.\frac{\partial \Pi}{\partial p_u}\right|_{m^*} = \frac{\partial \Pi}{\partial p_u} + \frac{\partial \Pi}{\partial p_L}\left[\left.\frac{dp_L}{dp_u}\right|_{m^*}\right]$$

Also, differentiating (10) directly along this path, at $p_u = 0$, and recognizing that N^L remains unchanged, gives

$$\text{(A14)} \quad \left.\frac{\partial \Pi}{\partial p_u}\right|_{m^*} = LR^T + N^L\left[\left.\frac{dp_L}{dp_u}\right|_{m^*}\right] + (p_S - c)\left[\left.\frac{\partial N^S}{\partial p_u}\right|_{m^*}\right]$$

Differentiating (6) and (7) and substituting these results and (A12) into (A14) yields

$$\text{(A15)} \quad \left.\frac{\partial \Pi}{\partial p_u}\right|_{m^*} = LR^T - N^L LR(m^*) + A$$

where

$$\text{(A16)} \quad A \equiv (p_S - c)\cdot \int_{m^*}^{\infty}\int_{p_S}^{\infty} h(B, p_S, m) f(m) dB dm$$

At the profit maximal p_S and p_L with $p_u = 0$, $\partial\Pi/\partial p_L = \partial\Pi/\partial p_S = 0$. Also, by hypothesis, $A > 0$ and $LR^T - N^L LR(m^*) > 0$. Thus, equating (A13) and (A15) shows that $\partial\Pi/\partial p_u > 0$

Now, consider increasing p_u from 0, holding p_L constant, and changing p_S at a rate $dp_S/dp_u < -LR^T/N^S$. Then, at $p_u = 0$, and with $\partial\Pi/\partial p_S = 0$,

$$d\Pi = \left(\frac{\partial \Pi}{\partial p_u}\right) dp_u + \left(\frac{\partial \Pi}{\partial p_S}\right) dp_S > 0$$

Also

$$dV = \left(\frac{\partial V}{\partial p_u}\right) dp_u + \left(\frac{\partial V}{\partial p_S}\right) dp_S$$

Calculating from (11), and using (1), (2), and (9),

$$\text{(A17)} \qquad \frac{\partial V}{\partial p_u} = -LR^T$$

So $dV = dp_u[-LR^T - N^S(dp_S/dp_u)] > 0$. Thus, there exist finite changes in p_u and p_S which increase both Π and V.

PROOF of Proposition 11:

As in the proof of Proposition 9, application of the envelope theorem gives

$$\text{(A18)} \quad dV^*/dp_u = d(V^* + \Pi^*)/dp_u = \partial V/\partial p_u + (\lambda + 1)\partial\Pi/\partial p_u$$

Combination of (A12), (A13), and (A15) yields at $p_u = 0$,

$$\text{(A19)} \quad \frac{\partial \Pi}{\partial p_u} = LR(m^*)\frac{\partial \Pi}{\partial p_L} + LR^T - N^L LR(m^*) + A$$

At the Ramsey optimum,

$$\text{(A20)} \qquad \frac{\partial V}{\partial p_L} + (\lambda + 1)\frac{\partial \Pi}{\partial p_L} = 0$$

Substituting $\partial V/\partial p_L = -N^L$, (A17), (A19), and (A20) into (A18) yields at $p_u = 0$,

$$\frac{dV^*}{dp_u} = \lambda[LR^T - N^L LR(m^*)] + (\lambda + 1)A$$

By the hypotheses, this is positive, and the proposition follows.

REFERENCES

Y. Barzel, "The Market for a Semipublic Good: The Case of the *American Economic Review*," *Amer. Econ. Rev.*, Sept. 1971, *61*, 665–74.

W. J. Baumol, "Scale Economies, Average Cost and the Profitability of Marginal Cost Pricing," in Ronald E. Grieson, ed., *Essays in Urban Economics and Public Finance in Honor of William J. Vickrey*, Massachusetts 1975.

______ **and D. F. Bradford**, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

______ **and Y. M. Braunstein**, "Empirical Study of Scale Economies and Produc-

tion Complementarity: The Case of Journal Publication," *J. Polit Econ.*, Oct. 1977, *85*, 1037–49.

S. Berg, "An Economic Analysis of the Demand for Scientific Journals," *J. Amer. Soc Info Sci*, Jan. 1972, *23*, 23–29.

M. Boiteux, "Sur la gestion des Monopoles Publics astréints à l'équilibre budgétaire," *Econometrica*, Jan. 1956, *24*, 22–40; translated by W. J. Baumol and D. F. Bradford as "On the Management of Public Monopolies Subject to Budgetary Constraints," *J. Econ. Theory*, Sept. 1971, *3*, 219 40.

Y. Braunstein, "Cost Data for Publication of Journals-Preliminary Analysis," disc paper no. 76-02, Center for Applied Economics, New York Univ. 1976.

——, "Economics of Journal Provision," unpublished paper, New York Univ. 1977.

Bernard Fry and Henry White, *Economics and Interaction of Publisher-Library Relationship in the Production and Use of Scholarly and Research Journals*, Final Report, NSF Grant GN-41398, Nov. 1975.

Donald W. Katzner, *Static Demand Theory*, New York 1970.

J. A. Ordover and R. D. Willig, "The Role of Information in Designing Social Policies Towards Externalities," econ. disc. paper no. 64, Bell Laboratories 1976.

—— **and** ——, (1977a) "On the Optimal Provision of Journals Qua Excludable Public Goods," econ. disc. paper no. 89, Bell Laboratories 1977.

—— **and** ——, (1977b) "Public, Private and Sometimes Shared Goods," mimeo., Bell Laboratories 1977.

J. C. Panzar and R. D. Willig, "Economies of Scale in Multioutput Production," *Quart. J. Econ.*, Aug. 1977, *91*, 481–94.

F. P. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47–61.

P. A. Samuelson, "The Pure Theory of Public Expenditures," *Rev. Econ. Statist.*, Nov. 1954, *36*, 381–89.

R. D. Willig and E. E. Bailey, "The Economic Gradient Method," mimeo, Bell Laboratories 1977.

U.S. Congress, Senate, Committee on the Judiciary, *Copyright Law Revision: Hearings on S 1361*, 93d Cong., July 31–Aug. 1, 1973.

Williams and Wilkins Co. v. The United States, 172 USPQ 670, 1972; 487 F.2d 1345, 1973; 420US 376, 1975.

Unfulfilled Long-Term Interest Rate Expectations and Changes in Business Fixed Investment

By JOHN C. WARNER*

The interest elasticity of investment is one of the oldest and most important unresolved issues in monetary theory. It is a key relationship in the most widely accepted theory of how Federal Reserve actions effect the economy.[1] Unfortunately, there is no consistent empirical verification of this relationship.[2] In spite of this lack of empirical support, the theoretical arguments for the interest elasticity of investment are so strong that most experts of monetary theory continue to write books and advise the government on the basis of this relationship. Yet policy exercises are suspect in the absence of some knowledge of the slope of the investment function. The purpose of this article is to use an error-learning model to investigate the modification of business investment decisions in response to errors in long-term interest rate expectations.

I. The Interest Expectations Error Hypothesis

Typically an investment project requires two years for its planning.[3] Sometime during the planning stage the project is analyzed for its economic feasibility. The test employed at this point dictates that the marginal efficiency of investment must exceed the long-term interest rate.[4] All of the projects that pass the test are allowed to continue in the planning phase. It is important to note that the firm has not made an irrevocable decision to invest. Although an investment project can be postponed at any stage of the planning and construction process, for this study it will be assumed that projects can be postponed easily before the date of financing whereas projects are postponed much more reluctantly after the date of financing. If a firm finances an investment project, this can be used as strong evidence that the firm is committed to the project. Mayer found that the completion of financing averages about three months prior to the start of construction (see his Table 4). Thus it is possible for marginal projects to be postponed immediately prior to the start of construction if the long-term interest rate is higher on the date of financing than the long-term interest rate used on the date of the preliminary screening. The interest rate effect on investment can be made symmetrical by assuming that some projects, for which the marginal efficiency of investment is below the long-term interest rate, will be continued in the planning phase. Thus it is also possible that investment projects might be reinstated if subsequent to the preliminary screening, and before the final decision to invest, the long-term interest rate decreases.[5]

The change in the interest rate between the preliminary screening and the final decision to invest can be represented by expression (1),

$$\frac{i_{actual} - i_{expected}}{i_{expected}} \tag{1}$$

which will be called the interest expectations error variable. It measures the percent error between the expected long-term interest rate on the screening date and the actual long-term interest rate on the date of financing. At the present time there is no empirical evidence available for measuring the time elapsing between the screening

*Assistant professor of economics, Stockton State College. This article is a summarization of my dissertation research conducted at West Virginia University under the direction of G. Richard Dreese.

[1] See Warren L. Smith.

[2] See Dale W. Jorgenson.

[3] See Thomas Mayer, Table 4.

[4] For a more detailed analysis of capital budgeting, see Ezra Solomon

[5] See William H. White, p. 278

date and the final decision to invest. Furthermore there is no evidence explaining how businessmen ascertain the interest rate which they expect to exist on the date of financing. The assumption was made that businessmen do not attempt to forecast the long-term interest rate and they just accept the actual long-term interest rate on the date of the preliminary screening as the credit markets best estimate of the future long-term interest rate.[6] By adopting this assumption, the expected long-term interest rate can be represented as the actual interest rate lagged x quarters from the date of financing. The subscript variable x is the time lag in quarters between the preliminary screening date and the final decision to invest date. Expression (1) can be rewritten as (2) by adopting the convention that i_{t-0} represents i_{actual} on the date of financing and i_{t-x} represents $i_{expected}$ on the date of the preliminary screening.

$$(2) \qquad \frac{i_{t-0} - i_{t-x}}{i_{t-x}}$$

By experimenting with different values for x it was determined that the best first approximation for the lag was five quarters.[7] Future research should investigate the possibility that the time lapse depends upon the type of investment project being undertaken. Ideally the time lapse should be represented as a weighted average of the time lapse for different investment projects.

The dependent variable, represented by expression (3), measures the percent error between the summation of the investment projects that pass the preliminary screening and the aggregate level of final decisions to invest.

$$(3) \qquad \frac{I_{final} - I_{preliminary}}{I_{preliminary}}$$

[6]See Burton G. Malkiel, pp. 217–18.

[7]One hundred and eight regressions of the complete model were run using different values for x and lags on the other variables in the model. With $x = 5$ the largest t-value was obtained for the coefficient for the interest expectations error variable. However, regardless of the value of x or the lags on the other variables, the coefficient for the interest expectations error variable easily passed the test for statistical significance.

As an initial effort to identify an existing data series that can be used as a proxy measure of final investment decisions (I_{final}), the unrealistic assumption was made that actual investment (I_{actual}) during a quarter is equal to the amount of final investment decisions (I_{final}) made during a quarter. At the theoretical level the two concepts are clearly not the same. That the dollar amounts are equal is correct only under the very restrictive condition that the backlog of capital appropriations does not change during the quarter. A more acceptable proxy measure of final investment decisions will be presented in the next section. The variable to be substituted at this point, I_{actual} (the amount of plant and equipment actually put in place during a quarter), should be understood to represent a *measure* of final investment decisions.

By allowing I_{final} to be represented by $I_{actual,t-0}$, $I_{preliminary}$ can be represented by $I_{expected,t-0}$, which is the result of an investment anticipation survey. Businessmen are asked to predict in period t − 2 the amount of plant and equipment that will be put in place during period t − 0. This investment anticipation survey should be understood to be a *proxy measure* of the preliminary decisions to invest during the quarter. The substitutions yield expression (4),

$$(4) \qquad \frac{I_{actual,t-0} - I_{expected,t-0}}{I_{expected,t-0}}$$

which has been labeled the investment realization ratio by Robert Eisner. Within the context of this article it will be designated the conventional ratio of final to preliminary investment decisions.

Expressions (2) and (4) are combined in equation (5) to specify the hypothesis that the conventional ratio of final to preliminary *investment decisions is a function of* the interest expectations error variable.

$$(5) \qquad \frac{I_{actual,t-0} - I_{expected,t-0}}{I_{expected,t-0}} = b_1 \frac{i_{t-0} - i_{t-5}}{i_{t-5}}$$

The hypothesis to be tested empirically is that the regression coefficient b_1 is negative and significantly different from zero. Ac-

cording to the marginal efficiency of investment schedule, if i_{t-0} is greater than i_{t-5}, $I_{actual,t-0}$ should be less than $I_{expected,t-0}$.[8] In other words, if the interest expectations error variable is positive, the ratio of final to preliminary investment decisions should be negative. The coefficient b_1 is a measure of the modifying effect that interest expectation errors have on business investment decisions.

II. The Improved Ratio of Final to Preliminary Investment Decisions

Equation (5) hypothesizes a behavioral relationship between the long-term interest rate and business investment decisions. The regression coefficient has been found to be not significant in previous studies for two reasons: 1) the construction of capital requires a period of time longer than one quarter, and 2) the capital-goods industry does not passively adjust the size of its output to changes in the rate of orders for new capital goods because of the associated adjustment costs. An article by Shirley Almon demonstrates that actual investment for any quarter is a weighted average of capital appropriations for the previous eight quarters. Thus actual investment represents a stage of commitment beyond the immediate control of the investing firms. The distributed-lag investment functions have made some progress in overcoming this obstacle, but there is one additional complication. The distributed-lag investment functions specify that the same proportion of investment projects are put in place during each quarter following the start of construction. The real world analogue of the fixed weight distributed-lag investment function is that the capital-goods producing industries exercise no control over their output, and that they automatically adjust output to any changes in the rate of new orders for capital goods. In the simplified world of elementary microeconomics adjustment costs can be ignored, but in the real world decisions to alter output involve considerable cost and require time. In fact the capital-goods industry does not mechanically adjust output to the rate of new orders. They allow the backlog of unfulfilled orders to serve as a buffer between the flow of new orders and the flow of shipments. From the investing firm's point of view, if the rate of new orders for any quarter exceeds the rate of actual investment, the difference will show up as an increase in the backlog of capital appropriations. The looseness of the connection between the new orders for capital and the actual output of the capital-goods industries has made the existence of a statistical relationship between the interest rate and investment virtually indetectable. This deficiency can be remedied on the theoretical level by incorporating into the neo-Keynesian transmission mechanism an explanation of how the capital-goods industries work down the backlog of unfilled orders.

In the light of the above discussion the focus of this study will be shifted to showing that at least the rate of business investment decisions is sensitive to the long-term interest rate. At the present, the best we can say is that in some loose and vague manner an increase in the new orders for capital goods is translated into an increase in the output of capital goods. Frank de Leeuw has demonstrated that actual investment plus the change in the backlog of capital appropriations during the quarter is an accurate measure of business investment decisions during the quarter. Expression (6) is called the improved ratio of final to preliminary investment decisions. It differs from expression (4) through the inclusion of the backlog of capital appropriations (BCA_{t-0}).

(6)

$$\frac{(I_{actual,t-0} - I_{expected,t-0}) + (BCA_{t-0} - BCA_{t-1})}{I_{expected\ t-0}}$$

Plugging the appropriate data into expressions (4) and (6) yields the conventional and improved ratios of final to preliminary investment decisions, respectively. Figure 1 dramatically illustrates that the volatility of business investment decisions are damped by the gradual output changes of the capi-

[8] See Edward Shapiro.

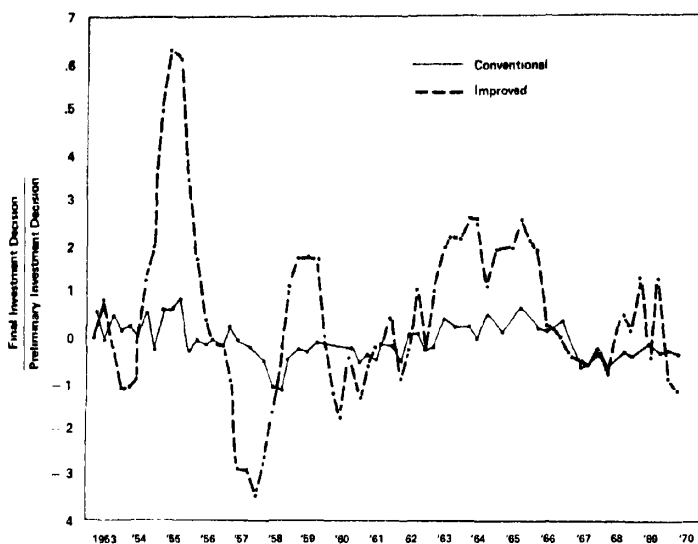



FIGURE 1 Conventional and Improved Ratios of Final to Preliminary Investment Decisions

tal-goods industries. By measuring investment instead of investment decisions, previous studies have eliminated much of the variation that economic theories of business investment behavior purport to explain. Upon reflection it is not hard to understand why an empirically validated relationship between interest and investment has been so difficult to establish.

As a brief exercise to reaffirm the efficacy of capital theory and to demonstrate the conceptual superiority of the improved ratio over the conventional ratio as an indicator of business investment decisions, a couple of important periods in the economic history of the United States have been chosen for analysis. In 1954 the Internal Revenue Code was revised and the excess corporate profits tax was repealed. The net result was a substantial reduction in the effective corporate profits tax rate. According to the present value formula, when the tax rate is reduced, more investment projects pass the test for economic feasibility. A comparison of the ratios for 1955 clearly shows that the increase in the flow of new orders greatly exceeded the ability of the capital-goods industry to keep pace. The conventional ratio reaches a peak of 8.6 percent in the fourth quarter of 1955. The improved ratio reaches a peak of 63.1 percent in the third quarter of 1955. During this time period sales were increasing and the interest rate was declining, but the substantial reduction in the tax that corporations expected to pay on profits must be a major factor in accounting for the robustness with which businesses modified their decisions to invest.

The second significant event in the recent economic history of the United States was the credit crunch of 1966. During 1966 the long-term interest rate rose by almost 100 basis points. The conventional ratio registered virtually no change whereas the improved ratio dropped from 20 percent in the first half of 1966 to 2 percent in the second half of 1966. The improved ratio indicates that the credit crunch was having a significant impact on the investment decisions of manufacturers. The gross movement of the improved ratio's time pattern is more compatable with economic intuition than the conventional ratio's time pattern. For the purpose of verifying the hypothesized behavioral relationship between the interest

rate and investment decisions, the improved ratio is clearly superior.

III. Additional Determinants of Investment: Sales and Taxes

Many studies of investment behavior have found that output is a significant determinant of investment.[9] The change in sales is the output variable chosen for this model. The symbol S in expression (7) represents quarterly sales. The subscript y represents the interval in quarters that businessmen use for making comparisons of sales. By substituting various values for y, it was determined that y equals four provided the highest t-value for the sales change variable. The four-quarter lag seems reasonable because changes in sales over a shorter time period are probably perceived by businessmen as the result of temporary or seasonal factors, whereas a four-quarter change in sales probably indicates a more permanent shift in necessary capacity.

$$(7) \qquad \frac{S_{t-0} - S_{t-y}}{S_{t-y}}$$

A third determinant of investment is the corporate income tax. A preliminary regression revealed that changes in the effective corporate income tax rate were not significantly correlated with the improve ratio. Allowance for the drastic change in the tax rate in 1954 was incorporated into the model by creating dummy variables for the second, third, and fourth quarters of 1955. The final form of the equation to be tested is presented in equation (8).

(8)

$$\frac{(I_{actual,t-0} - I_{expected,t-0}) + (BCA_{t-0} - BCA_{t-1})}{I_{expected,t-0}} = a + b_1 \frac{i_{t-0} - i_{t-5}}{i_{t-5}} + b_2 \frac{S_{t-0} - S_{t-4}}{S_{t-4}} + b_3 \mathrm{T} + e$$

T = 1; II, III, IV Quarters of 1955

T = 0; otherwise

The letter a stands for the constant and e is the stochastic term. The hypothesis to be tested with respect to the sales change and tax variables are that b_2 and b_3 are positive and significantly different from zero.

IV. Statistical Techniques and Data

The coefficients in equation (8) could be estimated by ordinary least squares regression but this would require the assumption that businessmen respond to errors in interest rate expectations and changes in sales within one quarter. The Almon lag technique was used to determine the length of the adjustment lag and the size of the weights within the lag period.[10] The procedure for determining the length of the lag is to specify a long lag and compute the weights. If any of the weights are negative, a shorter lag is specified by one-quarter increments until the negative weights disappear. This occurred for both the interest expectations error variable and the sales change variable at t − 3. The final form of the equation including the adjustment lag is presented in equation (9). A polynomial of degree four was used to generate the weights of the investment decision lag.

In preliminary trials of the model there was a significant amount of autocorrelation. Therefore the Cochrane-Orcutt iterative procedure was used to eliminate the possibility that autocorrelated disturbances might cause an underestimation of the sampling variance and thus deceptively increase the probability that the regression coefficients might pass the test for statistical significance.[11]

(9)

$$\frac{(I_{actual,t-0} - I_{expected,t-0}) + (BCA_{t-0} - BCA_{t-1})}{I_{expected,t-0}} = a + b_1 \sum_{m=0}^{3} w_m \frac{i_{t-m} - i_{t-5-m}}{i_{t-5-m}} +$$

[9] See Jorgenson.

[10] See Jan Kmenta, pp. 492 95.

[11] See J. Johnston, pp. 204–06.

$$+ b_2 \sum_{n=0}^{3} v_n \frac{S_{t-n} - S_{t-4-n}}{S_{t-4-n}} + b_3 \mathrm{T} + e$$

T = 1; II, III, IV Quarters of 1955

T = 0; otherwise

The data used to measure $I_{actual,t-0}$ were obtained from Genevieve Wimsatt and John Woodward.[12] The data series contained expenditures for nonmanufacturing as well as manufacturing industries. Since data on the backlog of capital appropriations are available only for manufacturing industries, the decision was made to include only manufacturing investment in the improved ratio of final to preliminary investment decisions.

The data for $I_{expected,t-0}$ were obtained from Wimsatt and Woodward.[13] The OBE-SEC survey asks businessmen to estimate the amount of investment that they expect to be completed two quarters in the future. The data reported in this survey are conceptually consistent with the actual manufacturing investment data reported in Wimsatt and Woodward ("Part I").

Due to the time necessary to produce capital and the institutional arrangements for making payments on acquired capital, there is a lag between the date that funds are appropriated for specific investment projects and the date that the funds are actually spent. The National Industrial Conference Board conducts a survey among manufacturing firms at the end of each quarter to determine the amount of funds that have been appropriated but not spent. These amounts are aggregated and published in the *Business Conditions Digest* (*BCD*) as the backlog of capital appropriations.[14] In order to convert this stock measure into a flow, the quarter-to-quarter changes in the backlog were calculated.

The data representing the long-term interest rate in equation (9) were obtained from the 1951 through 1970 issues of the *Survey of Current Business*. The raw data were Moody's composite index reported as a monthly average. For use in this model the monthly data were converted into quarterly averages.

The data for manufacturing sales were obtained from the *BCD*, p. 104. The insertion of a dummy variable to account for the tax change in 1954 was based upon an account of *The U.S. Economy in the 1950's* by Harold Vatter. The data available for all of the variables were sufficient to allow for sixty-four quarterly observations of the improved ratio of final to preliminary investment decisions. The time period covered by the study begins with the third quarter of 1954 and ends with the second quarter of 1970.

V. Empirical Results

The statistical results for equation (9) are reported in Table 1. The coefficient of determination adjusted for the degrees of freedom is .8505. This indicates that the proportion of variation in the improved ratio which is explained by the independent variables is substantial. The value of *rho* is .501. *Rho* is the coefficient of simple correlation between the disturbance term and the disturbance term lagged one quarter. The Chochrane-Orcutt iterative technique calculates rho in order to eliminate any first-order autocorrelation. The Durbin-Watson statistic fell within the range where we can accept the null hypothesis that there is no autocorrelation. The standard error of the estimate is .0734.

The critical t-value for a one-tail test with 58 degrees of freedom at the 5 percent level of significance is 1.68. The t-values for the coefficients b_1, b_2, and b_3 indicate that all of the independent variables are significantly correlated with the improved ratio. The coefficient b_1 which relates errors in long-term interest expectations to modifications of business investment decisions is -1.5487. The numerical interpretation is that if the interest rate increases 10 percent, manufacturers will decrease their decisions to invest by 15.5 percent. This estimate of the interest impact on investment decisions is much

[12] See their "Part I," Table 3, pp. 32–33.
[13] See their "Part II," Table 3, pp. 32–35.
[14] See *BCD*, p. 104.

TABLE 1—REGRESSION RESULTS FOR EQUATION (9)

Independent Variable	Coefficient	t-value	Weights[a]			
			t − 0	t − 1	t − 2	t − 3
Interest Expectations Error	−1.5487	−5 8968	15 9	17.3	28 5	38 3
Sales Change	2 0644	5 1598	36 6	23 1	16.6	23 7
Tax Change	2157	3 4675				
Constant	.0200	.6601				
R^2 = 8505						
$S E$ = .0734						
Rho = .501						
$D W$ = 2.30861						

[a]Shown in percent.

larger than any estimates reported in numerous previous studies. This can be explained by referring to two unique features of the model: 1) the dependent variable was designed to be sensitive to business investment decisions, and 2) a more realistic time pattern of the investment decision-making process was incorporated into the model. The weights of the distributed lag indicate that an error in interest expectations has 15.9 percent of its total impact on investment decisions in the current quarter and 17.3, 28.5, and 38.3 percent of its impact in the first, second, and third lagged quarters, respectively.[15]

The coefficient for the sales change variable is positive and statistically significant. The weights for the distributed lag indicate that a change in sales had 36.6 percent of its impact in the current quarter and 23.1, 16.6, and 23.7 percent of its impact in the first, second, and third quarters, respectively.

The dummy variable representing the change in the corporate income tax rate is also statistically significant. The value of the coefficient indicates that a large part of the surge in business investment decisions in the last three quarters of 1955 was attributable to the tax reduction [16]

The coefficient for the constant was not significant, thus we must accept the null hypothesis that the constant is zero. The fact that the constant is approximately zero supports the decision to use the OBE-SEC investment anticipation survey as the base upon which to calculate the improved ratio

On Figure 2 the actual and predicted

[15]In the light of the slow grinding process of formalized organizational decision making, it is not unrealistic to report that it can take up to three quarters for some companies to act on a decision after it has been made. See Glenn A Welsch The time is probably necessary to coordinate the capital expenditures budget with all of the other functional activities of the firm.

[16]The main reason for the dummy variable is that something extraordinary happened which caused investment decisions in the last three quarters of 1955 to increase more than can be explained by the interest and sales variables Using the present value formula as a guide to the potential determinants of investment and reviewing the economic history of that period it was concluded that the significant reduction in the effective corporate income tax was the only defensible explanation If the purpose of this article had been to analyze the impact of the corporate income tax on investment, this point would have been treated more extensively. The addition of the dummy variable improves the fit of the predicted ratios to the actual ratios but does not bias the regression coefficients on the interest or sales variables nor their statistical tests for significance Regardless of the reason for inserting the dummy variable if it was not appropriate the coefficient would not have been statistically significant. Exclusion of the dummy variable would have incorrectly attributed the dummy variables explanatory power to the two variables, interest and sales, which would have biased the coefficients, their statistical tests, and the lag structure.

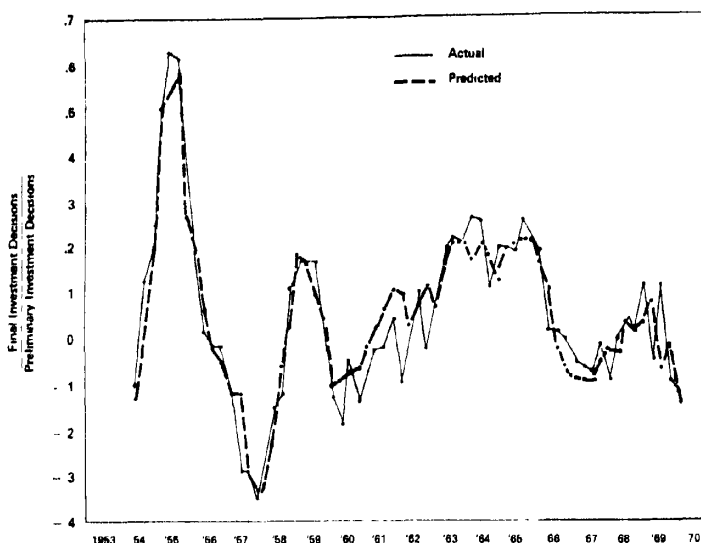



FIGURE 2 Improved Ratios of Final to Preliminary Investment Decisions

values of the improved ratio have been plotted. The close fit provides visual confirmation that the model has been correctly specified. In anticipation of the criticism that the model was tailored to fit the data, I attempted to predict the improved ratio for the period from the third quarter of 1970 to the first quarter of 1973. The standard error of the estimate for the predictions using extra-sample data was .0797. This does not differ substantially from the standard error of estimate (.0734), for the model using within-sample data.

The improved ratio derives most of its variability from the change in the backlog of capital appropriations. A simple test was constructed to determine if the quarter-to-quarter change in the backlog of capital appropriations could be explained by the interest expectations error hypothesis. The coefficient of simple correlation was calculated for the change in the backlog of capital appropriations, expression (10) and the interest expectations error variable, expression (11). The t-value was computed to be -6.4663, which is significant at the 5 percent level.

$$\frac{BCA_{t-0} - BCA_{t-1}}{BCA_{t-1}} \tag{10}$$

$$\frac{i_{t-3} - i_{t-8}}{i_{t-8}} \tag{11}$$

$$r = -.6199$$

VI. Conclusion

The fact that the capital goods industry allows the backlog of capital appropriations to act as a buffer between new orders for capital and the shipment or construction of capital substantially weakens the observable relationship between the changes in the long-term interest rate and changes in actual investment. In the absence of some intermediate mechanism which explains how the capital goods industry works down the backlog of orders, econometric tests of the interest elasticity of investment will be misleading. The model developed for this article bypasses the construction lag problem by including the change in the backlog of capital appropriations in the dependent variable. The second important innovation of the model is that explicit recognition is

given to the timing of the investment decision-making process. Received economic theory did not help much in this area. The appropriate lag structure was determined by a process of introspection and experimentation.[17]

The potential uses of the model are far from being exhausted. The interest effect on investment for nonmanufacturing sectors as well as numerous subcategories of the business sector remain to be investigated. Additional variables such as the tax rate, construction cost, and capacity utilization can be added to the model. Since capital is not a homogenous good, disaggregating capital according to type of capital or expected life should improve the accuracy of estimates of the interest effect on investment.

The results reported in this study demonstrate that the model is operational and that business investment decisions are sensitive to the long-term interest rate Before any policy implications can be drawn, however, the transmission mechanism of monetary policy will have to be studied in much greater detail. Further research along the lines pointed out in this article could substantially increase our knowledge of the channels through which Federal Reserve actions exert their impact upon the economy.

[17]This lag is attributable to the fact that the investment decision-making process is so complex that a substantial period of time is required for the professional staffs to analyze the numerous aspects of a project, see Welsch The investment decision-making process hypothesized in Section I reflects the business sectors most efficient tradeoff between the pressure to economize on professional staff and the necessity for corporations to effect a response to changing conditions in a competent and timely manner

REFERENCES

S. Almon, "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica*, Jan. 1965, *33*, 178-96.

F. de Leeuw, "The Demand for Capital Goods by Manufacturers: A Study of Quarterly Time Series," *Econometrica*, July 1962, *30*, 407-23.

R. Eisner, "Realization of Investment Anticipations," in James S. Duesenberry et al., eds, *The Brookings Quarterly Econometric Model of the United States*, Chicago 1965.

J. Johnston, *Econometric Methods*, New York 1963.

D. W. Jorgenson, "Econometric Studies of Investment Behavior: A Survey," *J. Econ. Lit.*, Dec. 1971, *9*, 1111-47.

Jan Kmenta, *Elements of Econometrics*, New York 1971.

B. G. Malkiel, "Expectations, Bond Prices, and the Term Structure of Interest Rates," *Quart J. Econ.*, May 1962, *76*, 197 218.

T. Mayer, "The Inflexibility of Monetary Policy," *Rev. Econ. Statist.*, Nov. 1958, *40*, 358-74.

Edward Shapiro, *Macroeconomic Analysis*, 2d ed., New York 1970

W. L. Smith, "A. Neo-Keynesian View of Monetary Policy," in *Controlling Monetary Aggregates*, Proc. Monetary Conference, Boston 1969.

Ezra Solomon, *The Management of Corporate Capital*, New York 1959.

Harold G. Vatter, *The U.S. Economy in the 1950's*, New York 1963.

Glenn A. Welsch, *Budgeting: Profit Planning and Control*, 3d ed., New Jersey 1971.

W. H. White, "The Timeliness of the Effects of Monetary Policy: The New Evidence from Econometric Models," *Banco Naz. Lavoro Quart. Rev.*, Sept. 1968, *21*, 276-303.

G. Wimsatt and J. T. Woodward, "Revised Estimates of New Plant and Equipment Expenditures in the United States 1947-1969: Part I," *Surv. Curr. Bus.*, Jan. 1970, *50*, 25-40; "Part II," Feb. 1970, *50*, 19-39.

Business Conditions Digest (*BCD*), Sept. 1971, *71*, 104.

Survey of Current Business, 1951-70.

Estimation of Complete Demand Systems from Household Budget Data: The Linear and Quadratic Expenditure Systems

By ROBERT A. POLLAK AND TERENCE J. WALES*

In this paper we explore two issues in empirical demand analysis the estimation of complete systems of demand equations using household budget data (Section I), and the incorporation of demographic characteristics into such systems (Section II). Although there are numerous studies which use time-series data to estimate complete demand systems, and many which estimate Engel curves from household budget data, there are virtually none which combine budget data from different periods to estimate complete systems. Since each budget study corresponds to a single price situation, one might conjecture that estimation of complete systems would not be possible unless budget studies were available from a large number of periods. This is incorrect. In Section I we argue that interesting complete demand systems can be estimated from a small number of budget studies, despite the limited price variability represented in such data. To demonstrate this, we use budget study data for two periods to estimate a pair of related demand systems the familiar linear expenditure system (*LES*) and a quadratic expenditure system (*QES*), a generalization of *LES* in which the demand equations are quadratic in total expenditure.[1]

The inclusion of demographic variables in the analysis of household consumption patterns is desirable because there are likely to be systematic differences in the consumption behavior of households with different demographic characteristics. But the effects of demographic characteristics have seldom been studied within the framework of complete demand systems. In Section II we propose "translating" as a general method for incorporating demographic variables into complete systems of demand equations. This method has not previously been discussed as a general technique for bringing demographic variables into demand systems, although it has been employed in several studies. To improve the estimates of the *LES* and the *QES* presented in Section I, we reestimate these two systems using translating to incorporate family size.

I. Estimation from Household Budget Data

In this section we discuss the use of household budget data from as few as two distinct price situations to estimate complete systems of demand equations. In particular, all of the $2n - 1$ independent parameters of the familiar *LES* are identified by two periods of household budget data, as are the $3n - 1$ independent parameters of the *QES* we estimate, a generalization of the *LES* in which the demand equations are quadratic in total expenditure. That is, both of these systems of demand equations are fully identified using budget study data from two periods, even though such observations correspond to only two sets of prices.

Household budget data record what each household in a sample spent on various

*University of Pennsylvania and University of British Columbia, respectively Pollak's research was funded in part by the National Science Foundation and Wales' research was funded in part by the Canada Council.

[1]To the best of our knowledge, there are only two previous studies which estimate complete demand systems from household budget data Kotaro Tsujimura and Tamotsu Sato use Japanese budget data for the period 1951-60 to estimate a dynamic version of the *LES*: they first estimate Engel curves separately for each year, and then use the coefficients to estimate the complete system for the entire period. Paul Deuster uses Indonesian data from two periods to estimate the *LES* We are grateful to A. S. Goldberger for bringing Deuster's work to our attention

consumption categories during a particular time period, as well as household income (perhaps broken down by source) and various demographic characteristics.[2] We assume that all households in a budget study for a particular period face the same prices.[3] A "quantity index" for each consumption category is obtained by dividing expenditure on the category by the corresponding price index. These quantity indexes are the empirical counterparts of the "goods" of traditional demand theory.

A complete system of demand equations describes the household's allocation of expenditure among some exhaustive set of consumption categories. It does not explain the division of "income" between "saving" and "consumption expenditure," nor the division of consumption expenditure between the included set of consumption categories and other consumption categories. A complete system of demand equations is said to be "theoretically plausible" if it is derivable from a "well-behaved" utility function, or, equivalently, if the demand equations are homogeneous of degree zero in prices and total expenditure, and the implied Slutsky matrix is symmetric and negative semidefinite. We are concerned with estimating complete systems of theoretically plausible demand equations from data which exhibit very little price variation, namely, household budget data from a small number of periods. When budget data are available for a large number of periods the estimation problems are essentially the same as with conventional time-series data.

We have ignored consumer durables in our analysis and focused on the allocation among a particular set of nondurable consumption categories of total expenditure on these categories. Underlying this procedure is the implicit assumption that nondurables are separable from the services of durables in the household's preference ordering. Consumption of the services of consumer durables can, under stringent assumptions (for example, perfect capital and second-hand markets), be disentangled from the full intertemporal allocation model and treated in the theoretical framework of one-period demand analysis. But even if we were willing to make the necessary assumptions, our data, which record purchases of durables but not stocks of durables held by households, would not support estimation.

Demographic characteristics such as family size are usually reported in household budget studies. We discuss the treatment of demographic variables in Section II; in this section we assume that the households in the sample do not differ in those demographic characteristics which affect tastes. Hence, random variations aside, all households in the sample have identical demand functions.[4]

For any demand system, household budget data for a single period identify the income-consumption curve corresponding to that period's prices, and each additional budget study identifies an additional income-consumption curve. Hence, two budget studies are sufficient to identify those systems of demand equations which are completely determined by two income-

[2]In practice published data often group households by income class and report the mean expenditure on each consumption category for each income class. We treat household budget data as if they reported separately the behavior of each household in the sample, in fn 16 we briefly discuss the complications of dealing with data which report only class means.

[3]If different households in the budget study face different prices (and if price data are available) then estimating complete demand systems using budget data is not very different from using time series. If, for example, the sample consists of households from different cities or regions, and if city or regional price data are available, then the fact that the data come from budget studies does not imply that price variation over the sample is severely limited. Examples of the use of regional price data in empirical demand analysis include James Dusenberry and Helen Kistin; Constantino Lluch; Lawrence Lau, Wuu-Long Lin, and Pan Yotopoulos. Further, even without regional price differences, the households in a budget study usually face different wage rates. Treating leisure as a good, Roger Betancourt used this fact to estimate a complete demand system (the *LES*) for Chile from data which exhibited no variation in the prices of consumption goods.

[4]The assumption of a homogeneous sample is not as restrictive as it might at first appear, since it is always possible and sometimes desirable to focus on homogeneous subsamples.

consumption curves. In general, whether a particular demand system can be estimated from T periods of household budget data depends on the number of goods as well as the general form of the demand equations. In systems like the translog (see Laurits Christensen, Dale Jorgenson, and Lau), or that derive from a quadratic direct utility function (see Leon Wegge), the number of parameters increases very rapidly with the number of goods, while in other systems, such as the *LES* or the *QES*, the number of parameters goes up less rapidly (linearly) with the number of goods.

The *LES* is a familiar system of demand equations which can be estimated from two periods of budget data. The demand equations corresponding to the *LES* (in expenditure form) are given by

$$(1)\quad p_i x_i = p_i h^i(P,\mu) = p_i b_i + a_i(\mu - \Sigma p_k b_k) \qquad \Sigma a_k = 1$$

where x_i denotes the quantity of the ith good, p_i its price and μ total expenditure on the n goods, hereafter referred to as "expenditure."[5] The system has $2n - 1$ independent parameters ($(n-1)a$'s and nb's) and is generated by the indirect utility function[6]

$$(2)\qquad \Psi(P,\mu) = -\frac{\Pi p_k^{a_k}}{\mu - \Sigma p_k b_k}$$

A household whose demand equations are of this form is often described as first purchasing "necessary," "subsistence," or "committed" quantities of each good (the b's) and then dividing the remaining or "superpumerary" expenditure among the goods in fixed proportions (the a's). In the *LES*, the "marginal budget shares"—the fractions of an additional dollar of expenditure spent on each good—are independent of prices and expenditure and are equal to the a's.[7]

In the *LES*, household budget data for a single period identify the a's; this follows from the fact that for any demand system such data identify the income-consumption curve corresponding to the period's prices, and hence the marginal budget shares at every level of expenditure. In the *LES*, the marginal budget shares are independent of the level of expenditure and are equal to the a's. If one of the b's is known a priori, then budget data for a single period is enough to identify not only the a's, but also the remaining $(n-1)b$'s.[8] Even if none of the b's are known a priori, budget data for two periods identify all of the parameters of the *LES*. Data from each period identify the corresponding income-consumption curve, and the intersection of the two linear income-consumption curves uniquely determines the point $(b_1, \ldots, b_n)$.

The *QES* (see Howe and the authors) is the class of demand systems which are quadratic in expenditure and which reduce to the *LES* for certain parameter values; marginal budget shares may vary with both

[5]Because we are dealing with household budget data which group households by income class, we distinguish between income and expenditure, where the latter is shorthand for "total expenditure on the consumption categories we are considering." Somewhat inconsistently, we defer to established usage and refer to "income-consumption curves" rather than "expenditure-consumption curves."

[6]We write the indirect utility function in this form to facilitate comparison with the *QES* we estimate, (4).

[7]The demand equations (1) are defined only in the region of the price-expenditure space for which they imply nonnegative consumption of every good. Regularity conditions require each x_i to be greater than the corresponding b_i, but they do not require the b's to be nonnegative (see Pollak, 1971). Demand is inelastic for positive b's and elastic for negative b's. The subsistence quantity interpretation of the b's breaks down when some of them are negative, and hence it is not appropriate to describe the *LES* in terms of subsistence quantities unless the b's are assumed to be nonnegative. We prefer to regard the signs of the b's as empirical questions, and we find that they are sometimes negative (see below and the Appendix).

[8]Howard Howe (1975) has shown that the identification of the parameters of Lluch's (1973) extended linear expenditure system can be interpreted in precisely these terms: Lluch's specification implicitly assumes that the b associated with saving is zero.

prices and expenditure.[9] In this paper we estimate the *QES* whose demand equations (in expenditure form) are given by

(3)

$$p_i x_i = p_i h^i(P,\mu) = p_i b_i + a_i(\mu - \Sigma p_k b_k) + (c_i - a_i)\lambda \Pi p_k^{-c_k}(\mu - \Sigma p_k b_k)^2$$

$$\Sigma a_k = 1, \quad \Sigma c_k = 1$$

Hereafter, we shall use the term *QES* to refer to this particular system, rather than the entire class. If $c_i = a_i$ for all i, then the quadratic terms vanish and the *QES* reduces to the *LES*. The *QES* has $3n - 1$ independent parameters ($(n-1)a$'s, $(n-1)$ c's, nb's, and one λ) and is generated by the indirect utility function

$$(4) \quad \Psi(P,\mu) = -\frac{\Pi p_k^{a_k}}{\mu - \Sigma p_k b_k} + \lambda \frac{\Pi p_k^{a_k}}{\Pi p_k^{c_k}}$$

The analogy with the *LES* suggests that the b's should not be described as subsistence quantities and $(\mu - \Sigma p_k b_k)$—the amount left over after purchasing the "necessary basket" as supernumerary expenditure, because this interpretation of the b's breaks down when some of them are negative

The marginal budget shares of the *QES* are given by

$$(5) \quad \frac{\partial}{\partial \mu} p_i h^i(P,\mu) = a_i + 2(c_i - a_i)\lambda \Pi p_k^{-c_k}(\mu - \Sigma p_k b_k)$$

They depend on all prices and expenditure, and on all of the parameters of the system (except in special cases). Unlike the *LES*, the relationship between the underlying parameters and the marginal budget shares is not a simple one, and no parameters of the system are identified by a single budget study.[10]

Analytic regularity conditions for the *QES* have not been derived, but it is possible to determine whether a particular *QES* satisfies the regularity conditions at a point in the price-expenditure space by evaluating the Slutsky matrix and observing whether it is negative semidefinite at that point.[11]

Budget data for two periods identify all of the parameters of the *QES*. The following heuristic argument, although not formally decisive, indicates why. Data from each period identify the income-consumption curve corresponding to that period's prices, and since all income-consumption curves radiate upward from the point $(b_1,\ldots,b_n)$, the intersection of two income-consumption curves determines the point $(b_1,\ldots,b_n)$. Estimates of the b's enable us to calculate supernumerary expenditure, $(\mu - \Sigma p_k b_k)$, for each household in each period. The a's are the coefficients of supernumerary expenditure, while the c's and λ can be disentangled from the coefficients of $(\mu - \Sigma p_k b_k)^2$.[12]

In contrast to the *LES* and *QES* which can be estimated from two cross sections, in the translog demand system the number of budget studies needed to estimate all of the parameters varies with the number of goods (n).[13] In particular, it can be shown that M cross sections provide enough information to permit estimation of a translog system with at most $n + 1 + M(n-1)$ parameters, while the nonhomothetic translog system contains $(n^2 + 3n - 2)/2$ in-

[9]Howe and the authors obtain a closed form characterization of the class of theoretically plausible demand systems which are quadratic in expenditure and estimate a *QES* using U.S. time-series data for the postwar period.

[10]But if the marginal budget shares are independent of expenditure, the system reduces to the *LES* and the a's are identified by a single budget study

[11]A straightforward continuity argument shows that the *QES* is a nontrivial generalization of the *LES*. If the a's and b's are such that the associated *LES* satisfies the regularity conditions at a point in the price-expenditure space, then, for values of the c's sufficiently close to the corresponding a's, the *QES* also satisfies the regularity conditions at that point

[12]More formally, suppose that we rewrite (3) as $p_i x_i = \theta_{1i} + \theta_{2i}\mu + \theta_{3i}\mu^2$ and estimate the θ's for $n-1$ goods in both periods. Then it is easy to show that after substituting the θ_{3i} values into the θ_{2i} values we can obtain estimates of the a's, and of $\Sigma p_k b_k$ for both periods. Using these we can then obtain estimates of the individual b's from the θ_{1i} values, and estimates of the c's and λ from the θ_{3i} values

[13]See Lau, Lin, and Yotopoulos for estimates of a translog for Taiwan based on household budget data.

dependent parameters. Hence, two cross sections are sufficient to identify the translog with two or three goods, but not the translog with four or more goods; three cross sections identify it with four goods, but not with five or more.

We have estimated the *LES* and *QES* using data from two budget studies. The data are reported in the *Family Expenditure Survey* series, an annual publication which reports income and expenditure patterns of households in the United Kingdom. We use data for the two years 1966 and 1972 (see Department of Employment and Productivity 1966, Tables 5-7; 1972, Tables 11-13), since this represents the longest interval over which the family size characteristics are defined consistently.[14] To simplify the computations, we have analyzed only three consumption categories, "food," "clothing," and "miscellaneous."[15] Current expenditures on these categories were obtained directly from the surveys, and the corresponding price indexes from *National Income and Expenditure 1964-74* (see Central Statistical Office, Tables 29, 30).

The *Family Expenditure Survey* cross classifies households by income and family size. For 1966, it reports mean expenditure on each consumption category by families in four income classes and three family size classes: "one child," "two children," and "three or more children." For 1972, we use the six income classes for families with one child and two children, five income classes for families with three children, and three income classes for families with "more than three children." The survey reports mean consumption expenditure on each category for each of these thirty-two cells, and we have treated these as our basic data. Furthermore, we have treated them as if they represented the consumption patterns of households rather than cell means.[16]

We obtain a stochastic form for the *LES* and *QES* by adding a disturbance term to the share form of each demand equation, where the share forms are obtained by dividing the expenditure forms (1) and (3) by expenditure. We used the share forms because they are likely to involve less heteroskedasticity than the expenditure forms. We denote the 3×1 vector of disturbances corresponding to the ith cell by $e_i = (e_{i1}, e_{i2}, e_{i3})$ and assume that $E(e_i) = 0$, that $E(e_i e_i') = \Omega$ for all i, and that the e_i are independently normally distributed. Since the dependent variables and the nonstochastic terms in the equations are shares and add to one for each cell, the covariance matrix is singular. Hence maximum likelihood estimates of the system can be obtained by minimizing the determinant of the sample error covariance matrix with respect to the parameters after dropping any equation.

Estimates of the *LES* and *QES* parameters based on the data just described are

[14]We discuss this in Section II, fn. 25.

[15]Our food category does not include alcoholic drink. Our clothing is officially "clothing and footwear." Our miscellaneous is the sum of two categories from the survey, "other goods" and "services." The principal subcategories of other goods are "leather, travel and sports goods, jewelry, fancy goods, etc.," "books, magazines and periodicals," "toys and stationery goods, etc.," and "matches, soap, cleaning materials, etc." The principal subcategories of services are "radio and television, licenses and rentals," "educational and training expenses," and "subscriptions and donations; hotel and holiday expenses; miscellaneous other services." The survey reports seven major expenditure categories which we have omitted entirely "housing," "fuel, light, and power," "alcoholic drink," "tobacco," "durable household goods," "transport and vehicles," and "miscellaneous." Our three categories of food, clothing, and miscellaneous account for between 46 and 58 percent of total consumption expenditures.

[16]Since our consumption data are cell means rather than observations on individual families, an aggregation problem arises. With the *LES*, it causes no difficulty because the mean consumption pattern of a group is the consumption pattern corresponding to the group's mean expenditure. But with the *QES*, the mean consumption pattern of a group depends on the variance as well as the mean of the group's expenditure. Unfortunately, we do not know these variances, so, *faute de mieux*, we have treated the reported cell means as if they were observations on individual families. Aggregation of demand systems quadratic in the independent variable is discussed by Lawrence Klein, pp. 25-26, and W. E. Diewert, pp. 129-30.

TABLE 1—MARGINAL BUDGET SHARES AND OWN-PRICE ELASTICITIES FOR THE *LES* AND THE *QES* AT 1972 PRICES WITH TASTES INDEPENDENT OF FAMILY SIZE

Expenditure (S/Week)	*QES* Marginal Budget Shares			*QES* Own-Price Elasticities			*LES* Own-Price Elasticities		
	Food	Clothing	Misc.	Food	Clothing	Misc.	Food	Clothing	Misc.
225	.46	.26	.28	−.42	−.10	−.23	−.51	−.75	−1.02
325	.40	.23	.37	−.54	−.54	−.77	−.61	−.85	−1.00
425	.34	.20	.46	−.71	−.79	−1.53	−.67	−.89	−1.00
525	.29	.17	.54	−.93	−.99	−2.24	−.72	−.91	−1.00
625	.23	.14	.63	−1.19	−1.20	−2.86	−.75	−.93	−1.00
725	.17	.11	.72	−1.49	−1.43	−3.39	−.78	−.94	−1.00

Note: The sample expenditure range in 1972 is 232–684 S/week. For the *LES* the corresponding marginal budget shares are independent of expenditure and are .35, .20, and .45.

presented in the Appendix. These estimates are based on the assumption that the thirty-two "families" have identical tastes, an assumption we drop in Section II when we discuss the treatment of demographic variables. The Slutsky matrix implied by our parameter estimates is negative semidefinite at each of the thirty-two price-expenditure situations corresponding to our data with the *LES* and for all but two with the *QES*.[17]

Perhaps the most interesting comparison to be made between the *LES* and the *QES* involves the behavior of the marginal budget shares. For the *LES* they are the estimated a_i parameters, while for the *QES* they are given by the expression in equation (5), and depend on all of the estimated parameters. In Table 1 we present the marginal budget shares corresponding to 1972 prices implied by our parameter estimates for six expenditure levels that encompass the 1972 sample expenditure levels.[18] For the *QES* the shares vary considerably with expenditure levels: 46 percent of an additional shilling goes to food at the lowest level, and only 17 percent at the highest; for the *LES* the share is constrained to be the same at all expenditure levels and is 35 percent. Corresponding patterns for clothing and miscellaneous can be read from the table. The own-price elasticities are also presented in Table 1, and lie between −.1 and −3.39 for the *QES* and −.51 and −1.02 for the *LES*. The *QES* own-price elasticities vary substantially more with expenditure levels than those implied by the *LES*, and the two sets of elasticity estimates suggest different patterns of responsiveness to price changes.

The *QES* represents a significant improvement over the *LES* according to the usual likelihood ratio test. The value of $-2 \ln \omega$, where ω is the ratio of the likelihoods associated with the *LES* and the *QES*, is 8.2, which is significant at the 5 percent level.

II. Demographic Characteristics

Demographic variables have traditionally played a major role in the analysis of household budget data. Instead of assuming, as we did in Section I, that all households in the sample have identical tastes, only those with the same demographic profiles are assumed to have the same demand functions. Family size and composition, race and religion, age and education have all been

[17]Our approach to the evaluation of regularity conditions differs from that adopted by Jorgenson and Lau in their discussion of the translog. They consider only whether regularity conditions are satisfied at a single point, the point of approximation of the translog, whereas we have checked them at every sample point. On the other hand, Jorgenson and Lau test the significance of the ability of their estimated translog to satisfy the regularity conditions, whereas we do not.

[18]Since the results for 1966 are similar we have not presented them. However, the marginal budget shares and elasticities corresponding to any price-expenditure situation can easily be calculated from the parameter estimates reported in the Appendix.

used as demographic variables in demand studies, although seldom in the context of complete systems of demand equations.[19]

There are two ways to proceed. First, without any additional assumptions, the methods of Section I can be applied to subsamples of households with identical demographic profiles. For example, we might specify that each household's demand equations are given by a *QES*, and estimate the $3n - 1$ independent parameters of that system separately for each household type. This approach allows all of the parameters of the demand system to depend on the demographic variables, and does not require us to specify the form of the relationship between these parameters and the demographic variables. Under this approach, the only data relevant to the analysis of households with a particular demographic profile are observations on households with that profile.

The second approach introduces assumptions which relate the behavior of households with different demographic profiles. For example, we shall assume that each household's demand equations are given by a *QES*, that the b's depend linearly on some set of demographic variables, and that the remaining parameters are independent of the demographic variables. Under this set of assumptions we can draw inferences about households with one demographic profile from observations on the behavior of households with different profiles, a type of inference which is not possible under the first approach. Our empirical work deals only with the number of children in the family, but the techniques we use are readily applicable to other demographic characteristics.

We now describe translating, a general procedure for incorporating demographic variables into demand systems.[20] Under this procedure the original demand system is first replaced by a new system which contains parameters suitable for introducing such variables, and it is then assumed that these newly introduced parameters are the only ones which depend on the demographic variables. The process is completed by specifying the functional form relating these parameters to the demographic variables. More specifically, translating replaces the original demand system $x_i = \bar{h}^i(P,\mu)$ by the modified system

$$(6) \quad h^i(P,\mu) = d_i + \bar{h}^i(P,\mu - \Sigma p_k d_k)$$

and specifies that only the d's depend on demographic variables. It is tempting to interpret the "translation parameters" as "necessary quantities," but this can be misleading.[21]

If translation parameters are already present in the original system of demand equations (as they are in the *LES* and the *QES*), then the translation approach reduces to a specification of which parameters of the original system depend on demographic variables and which do not. If the original system is generated by a Cobb Douglas or *CES* utility function, then introducing translation parameters yields the *LES* or the "generalized *CES*" (see Pollak, 1970 or 1971; Wales). If the original system is translog, then translating yields the "translog with committed quantities" (see Marilyn Manser). In general, if the original

[19]The classic reference is S. J. Prais and Hendrik Houthakker. Recent examples include Howe (1974); Lau, Lin and Yotopoulos; Marjorie McElroy Richard Parks and Anton Barten use cross-country differences in age composition to analyze aggregate time-series data from eleven countries.

[20]An alternative procedure is proposed by Barten, whose discussion is amplified and corrected by John Muellbauer. Alan Brown and Angus Deaton, pp. 1178-86, survey the more traditional literature on "equivalent adult scales" and supply references to the literature.

[21]There are two distinct reasons. First, the d's can be negative. Second, with some demand systems such as that implied by the translated constant elasticity of substitution (*CES*) direct utility function, $U(X) = -\Sigma a_k(d_k - x_k)^c$, for $c > 1$, regularity conditions require d_i to be greater than x_i; the additive quadratic direct utility function discussed by Paul Samuelson, p. 93, is a member of this class (see Pollak, 1971). In these cases, it is natural to interpret the d's as "bliss points."

system was generated by the indirect utility function $\overline{\Psi}(P,\mu)$, then it is easy to verify that the modified system is generated by

$$\Psi(P,\mu) = \overline{\Psi}(P,\mu - \Sigma p_k d_k) \quad (7)$$

When translating is used to introduce demographic characteristics into complete demand systems, it is easy to verify that there is a close relation between the effects of changes in demographic variables and the effects of changes in total expenditure. A change in any demographic variable (for definiteness, an increase in family size) causes a reallocation of expenditure among the consumption categories. That is, since total consumption expenditure remains unchanged, increases in the consumption of some goods must be balanced by decreases in the consumption of others. Hence, the sign of the effect on $p_i x_i$ of a change in family size cannot be inferred from the sign of its effect on d_i.[22] Changes in the demographic variables affect all of the d's simultaneously, so, for example, we cannot infer an increase in $p_i x_i$ from an increase in family size which increases d_i. Furthermore, there is no presumption that an increase in family size will increase rather than decrease the d's: in either case, changes in the d's imply a reallocation of total expenditure among the goods which leaves total expenditure unchanged.

Demographic effects can be introduced into any system of demand equations by allowing any subset of parameters to depend on demographic variables. Translating postulates that demographic effects operate through a particular subset of n independent parameters which enter the demand system in a relatively simple way. In principle, the question of which parameters depend on demographic variables is an empirical one, and the specification implied by translating can be tested against the more general hypothesis that additional parameters also depend on demographic variables.[23]

Using translating to introduce family size into the *LES* and the *QES*, we reestimate these systems with the *Family Expenditure Survey* data described in Section I. Since the *Family Expenditure Survey* reports the consumption patterns of households containing two adults (one man and one woman) and a specified number of children, the data are better suited to studying the effect of the number of children on household consumption patterns than data which report only household size.[24] We use the two years 1966 and 1972 because this choice corresponds to the longest span over which "children" were defined consistently.[25]

To incorporate the effect of the number of children into the *LES* and the *QES*, we assume that the b's of these systems depend linearly on the number of persons in the household, N:[26]

$$b_i = b_i^* + \beta_i N \quad (8)$$

In the *LES*, there are now $3n - 1$ parameters to be estimated (nb^*'s, $n\beta$'s, and $n - 1$ a's); in the *QES* there are $4n - 1$ param-

[22]On the other hand, with habit formation a change in past consumption of x_i affects d_i but leaves the other d's unchanged (see Pollak, 1970), so the effect of a change in the past consumption of a particular good on its current consumption can be inferred.

[23]The treatment of demographic variables in Lau, Lin, and Yotopoulus is equivalent to a specification in which $n - 1$ independent parameters of the homogeneous translog depend linearly on the logarithms of the demographic variables, although they do not present their specification in this way

[24]For example, published tables from the *BLS Study of Consumer Expenditures 1950* and *Survey of Consumer Expenditures 1960-61* cross classify households by income and the number of individuals in the household.

[25]In the surveys before 1966, the number of persons in the household rather than the number of children was reported. After 1972, persons were classified as children if they were 18 or under, while in earlier years they were so classified if they were 16 or under

[26]The 1966 survey reports the average number of persons per household for each income class for households consisting of "one man, one woman, and three or more children," as does the 1972 survey for households consisting of "one man, one woman, and more than three children." In these cases, we have used average number of persons per household for N.

TABLE 2 MARGINAL AND AVERAGE BUDGET SHARES FOR THE *LES* AND THE *QES* AT 1972 PRICES WITH TASTES DEPENDENT ON FAMILY SIZE

	Food			Clothing			Miscellaneous		
	Children			Children			Children		
Expenditure *S*/week	1	2	3	1	2	3	1	2	3
				Marginal Budget Shares—*QES*					
225	37	40	43	.26	28	30	.37	.32	.27
325	31	34	37	22	24	26	.47	.42	.37
425	.24	27	30	.18	20	.22	58	53	.48
525	.18	21	24	.13	15	17	69	.64	.59
625	.11	14	17	09	11	.13	.80	.75	.70
725	.05	.08	11	05	07	.09	.90	.85	.80
				Average Budget Shares—*QES*					
225	.61	65	69	14	.12	.09	25	23	22
325	53	57	.60	17	16	15	.30	.27	25
425	47	50	54	18	.18	.17	35	32	.29
525	42	45	.49	17	18	18	41	.37	.33
625	38	41	44	16	.17	.17	.46	.42	.39
725	33	37	40	15	16	.16	.52	.47	44
				Average Budget Shares—*LES*					
225	62	68	.73	14	13	13	24	19	.14
325	52	56	60	16	16	15	32	.28	25
425	.47	50	53	17	17	17	36	33	.30
525	44	46	48	18	18	18	38	.36	34
625	41	43	45	19	18	.18	40	.39	37
725	40	41	43	19	19	.19	41	40	38

Note The sample expenditure range in 1972 is 232 684 *S*/week For the *LES*, marginal budget shares are independent of expenditure and family size and are 30, 20, and 49

eters (nb^*'s, $n\beta$'s, $(n-1)a$'s, $(n-1)c$'s, and λ). For both of these systems, this is equivalent to incorporating these demographic variables by translating.

The stochastic structure is obtained by adding a disturbance term to the share form of each nonstochastic demand equation, just as in Section I. The parameter estimates obtained for the *LES* and *QES* are presented in the Appendix. The regularity conditions are satisfied (i.e., the implied Slutsky matrix is negative semidefinite) for all but four of the thirty-two price-expenditure situations corresponding to families in the sample with the *LES*, and for all but one with the *QES*.

Once again the *QES* represents a significant improvement over the *LES* with the likelihood ratio test statistic taking a value of 17.7 while the 1 percent significance level is 11.3. Further, inclusion of the three additional family size parameters is highly significant for both functional forms, with likelihood ratio test statistics of 66.7 and 76.3 for the *LES* and *QES*, respectively, while the 1 percent significance level is 11.3.

The marginal budget shares for the *QES* depend on prices, expenditure, and family size; in Table 2 we report the implied marginal budget shares for families with one, two, and three children at 1972 prices for the expenditure levels reported in Table 1.[27] As might be expected, the marginal budget shares increase with family size (expenditure held constant) for food and clothing, and decrease for miscellaneous. The pattern with respect to changes in expenditure levels (family size held constant) is similar to that reported and discussed in Section I.

[27]The marginal budget shares at 1966 prices were very similar.

For the *LES*, when family size is incorporated through the *b*'s, the marginal budget shares are independent of prices, expenditure, and family size. Our estimates of the marginal budget shares are .30, .21, and .49 for food, clothing, and miscellaneous, respectively.[28]

An increase in family size (with total expenditure held constant) causes a reallocation of expenditure which is reflected by changes in the average budget shares; the average budget shares implied by our parameter estimates are presented in Table 2 for both the *LES* and the *QES*. For both systems and all expenditure levels, the average budget share of food increases with family size, while the share of miscellaneous decreases, implying a reallocation from miscellaneous to food. For the *QES*, at low expenditure levels the average budget share for clothing falls as family size increases, but at higher expenditure levels it is virtually independent of family size; for the *LES* it is independent of family size at all expenditure levels.

The pattern of average budget shares for the *LES* is fairly close to that for the *QES* (particularly in the middle expenditure ranges). The estimated b_i values ($b_i = b_i^* + \beta_i N$) are all positive for the *QES*, but are negative for large family sizes for clothing and miscellaneous using the *LES* (see the Appendix) suggesting that it is inappropriate to interpret the *b*'s as necessary or subsistence quantities. In the *QES* the *b*'s increase with family size, while in the *LES* an increase in N implies a decrease in the *b*'s.[29] Finally, it should be noted that the *QES* average budget shares for clothing do not vary monotonically with expenditure (for fixed family size); in the *LES*, average budget shares must either increase monotonically or decrease monotonically with expenditure.

III. Conclusion

We have estimated two complete systems of demand equations—the *LES* and the *QES*—using *U.K.* household budget data for two periods, and investigated the influence of family size on consumption patterns within the framework of these systems. The empirical results were generally good, regularity conditions were almost always satisfied, and our results improved significantly as we moved from the *LES* with no family size effects to the *QES* with family size effects.

The *QES* is a new generalization of the *LES* which is particularly well suited to the analysis of data sets which exhibit severely limited price variation. With household budget data, each study identifies an income-consumption curve, and, although none of the underlying parameters of the *QES* can be identified from a single income-consumption curve, all of the parameters can be identified from two such curves. The effect of demographic variables on consumption patterns has traditionally been analyzed using household budget data, but seldom in the context of complete demand systems. Complete demand systems have two principal advantages over other frameworks for analyzing the effects of demographic variables. First, by incorporating the budget constraint into the analysis, the complete system approach forces recognition of the fact that an increase in expenditure on one consumption category must be balanced by decreases in the expenditure on others. Second, the complete system approach permits the separation of demographic effects from own- and cross-price effects as well as from income effects. Unless such a separation is made, there is no presumption that demographic effects estimated from one price situation will be relevant in another. We incorporated family size into the *LES* and the *QES* by allowing a particular subset of their parameters (the *b*'s) to depend linearly on family size.

[28] The comparable marginal budget shares obtained in Section I for the *LES* were .35, .20, and .45

[29] This does not imply that larger families can achieve a particular standard of living with less expenditure than small families. Empirically determined equivalence scales play an important role in demand analysis, but there is no justification for using them to make welfare comparisons.

The techniques we have used in this paper have implications for empirical demand analysis which go well beyond the particular functional forms we have estimated and the demographic variable we have employed. First, translating, the method we employed to incorporate family size into the *LES* and the *QES* is a technique which can be used to incorporate any set of demographic variables into any complete system of demand equations. Under translating, demographic effects operate through a particular set of n independent "translation parameters" which happen to be present in both the *LES* and the *QES*. But if these parameters are not present in a particular demand system, they can be introduced in a straightforward way.

Second, it has generally been assumed that complete demand systems could not be estimated from household budget data unless such data are available for many periods. We have demonstrated that interesting complete systems of demand equations can be estimated on the basis of household budget data which exhibit very little price variation. As budget study data become increasingly available, estimation of demand systems using such data from a limited number of periods will become an increasingly important technique for empirical demand analysis

APPENDIX

A Estimated *LES* Equations

Parameter	*LES*		*LES*	
a_1	.351	(11 5)	297	(24 6)
a_2	202	(13 7)	213	(15 8)
a_3	.447	(15 7)	490	(26 9)
b_1^*	75 7	(2.2)	182 9	(2 1)
b_2^*	7.3	(4)	113 3	(1 8)
b_3^*	− 88	(02)	225 3	(1 6)
β_1			−23 1	(1.1)
β_2			−25 4	(1 7)
β_3			−54.8	(1.7)
c_1				
c_2				
c_3				
λ				
R_F^2	.572		940	
R_C^2	186		.385	
R_M^2	.503		.820	

B. Estimated *QES* Equations

Parameter	*QES*		*QES*	
a_1	442	(5 9)	.409	(6.6)
a_2	.253	(7.9)	.289	(5.8)
a_3	305	(2.3)	.302	(5.1)
b_1^*	104 5	(6 3)	23.6	(.8)
b_2^*	24.8	(2.6)	−6.5	(.3)
b_3^*	39.6	(3.5)	2.3	(.1)
β_1			18 4	(7.0)
β_2			5 9	(3.0)
β_3			7 1	(2.3)
c_1	−4 263	(1 4)	−.542	(4)
c_2	−2 325	(1 0)	− 347	(3)
c_3	7.588	(2 3)	1 889	(3.3)
λ	00019	(2.4)	.000593	(8)
R_F^2	602		.963	
R_C^2	226		377	
R_M^2	598		888	

Notes Data are measured in shillings per week.

The normalization rules used are $\sum_1^3 a_k = 1$ and $\sum_1^3 c_k = 1$

Food is the first, clothing the second, and miscellaneous the third good.

The ith share equation for the *QES* is given by

$$w_i = \frac{p_i}{\mu}(b_i^* + \beta_i N) + a_i\left(1 - \sum \frac{p_k}{\mu}(b_k^* + \beta_k N)\right) + (c_i - a_i)\lambda \Pi \left(\frac{p_k}{\mu}\right)^{-c_k}\left(1 - \sum \frac{p_k}{\mu}(b_k^* + \beta_k N)\right)^2$$

For the cases in which family size does not affect demand the β's are set to zero

For the *LES* $c_i = a_i$ for all goods.

The numbers in parentheses are ratios of parameters to asymptotic standard errors

REFERENCES

A. P. **Barten,** "Family Composition, Prices and Expenditure Patterns," in P. Hart et al., eds., *Econometric Analysis for National Economic Planning: 16th Symposium of the Colston Society*, London 1964, 277 92.

R. R. Betancourt, "The Estimation of Price Elasticities from Cross-Section Data Under Additive Preferences," *Int. Econ. Rev.*, June 1971, *12*, 283-92.

A. Brown and A. Deaton, "Surveys in Applied Economics: Models of Consumer Behaviour," *Econ. J.*, Dec. 1972, *82*, 1145-236.

L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Utility

Functions," *Amer. Econ. Rev.*, June 1975, *65*, 367–83.

P. R. Deuster, "Rural Consequences of Indonesian Inflation: Case Study of the Joghakarta Region," unpublished doctoral dissertation, Univ. Wisconsin 1971.

W. E. Diewert, "Applications of Duality Theory," in Michael D. Intriligator and David A. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam 1974, 106–71.

J. S. Duesenberry and H. Kistin, "The Role of Demand in Economic Structure," in Wassily Leontief et al., eds., *Studies in the Structure of the American Economy*, New York 1953, 451–82.

H. Howe, "Estimation of the Linear and Quadratic Expenditure Systems: A Cross-Section Case for Columbia," unpublished doctoral dissertation, Univ. Pennsylvania 1974.

———, "Development of the Extended Linear Expenditure System from Simple Saving Assumptions," *Euro. Econ. Rev.*, July 1975, *6*, 304–10.

———, **R. A. Pollak, and T. J. Wales**, "Theory and Time Series Estimation of the Quadratic Expenditure System," disc. paper no. 388, Univ. Pennsylvania, Oct. 1977.

D. W. Jorgenson and L. J. Lau, "The Integrability of Consumer Demand Funtions," Harvard Instit. Econ. Res., disc. paper no. 425, Harvard Univ., July 1975.

Lawrence R. Klein, *An Introduction to Econometrics*, Englewood Cliffs 1962.

L. J. Lau, W. L. Lin, and P. A. Yotopoulos, "The Linear Logarithmic Expenditure System: An Application to Consumption-Leisure Choice," memo. no. 187, Center Res. Econ. Growth, Stanford Univ., Apr. 1975.

C. Lluch, "Consumer Demand Functions, Spain, 1958–1964," *Euro. Econ. Rev.*, Spring 1971, *2*, 277–302.

———, "The Extended Linear Expenditure System," *Euro. Econ. Rev.*, Apr. 1973, *4*, 21–31.

M. B. McElroy, "A Spliced *CES* Expenditure System," *Int. Econ. Rev.*, Oct. 1975, *16*, 765–80.

M. E. Manser, "The Translog Utility Function with Changing Tastes," Bur. Labor Statist. work. paper no. 33, Office of Research Methods and Standards, Jan. 1975.

J. Muellbauer, "Household Composition, Engel Curves and Welfare Comparisons between Households," *Euro. Econ. Rev.*, Aug. 1974, *5*, 103–22.

R. W. Parks and A. P. Barten, "A Cross-Country Comparison of the Effects of Prices, Income and Population Composition on Consumption Patterns," *Econ. J.*, Sept. 1973, *83*, 834–52.

R. A. Pollak, "Habit Formation and Dynamic Demand Functions," *J. Polit. Econ.*, July/Aug. 1970, *78*, Part 1, 745–63.

———, "Additive Utility Functions and Linear Engel Curves," *Rev. Econ. Stud.*, Oct. 1971, *38*, 401–14.

S. J. Prais and Hendrik S. Houthakker, *The Analysis of Family Budgets*, Cambridge 1955.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

K. Tsujimura and T. Sato, "Irreversibility of Consumer Behavior in Terms of Numerical Preference Fields," *Rev. Econ. Statist.*, Aug. 1964, *46*, 305–19.

T. J. Wales, "A Generalized Linear Expenditure Model of the Demand for Non-Durable Goods in Canada," *Can. J. Econ.*, Nov. 1971, *4*, 471–84.

L. L. Wegge, "The Demand Curves from a Quadratic Utility Indicator," *Rev. Econ. Stud.*, Apr. 1968, *35*, 209–24.

Central Statistical Office, *National Income and Expenditure 1964–74*, London 1975.

Department of Employment and Productivity, *Family Expenditure Survey*, London 1966; 1972.

U.S. Bureau of Labor Statistics, *Survey of Consumer Expenditures 1960–61*, "Consumer Expenditures and Income, Cross-Classification of Family Characteristics," Suppl. 2 to *BLS* Report 237–93, Washington, June 1966.

———, *Study of Consumer Expenditures, Income, and Savings*, 1950 (Univ. of Pennsylvania 1957).

Why Women Earn Less: The Theory and Estimation of Differential Overqualification

By ROBERT H. FRANK*

The average wage rate for females in the United States has been approximately two-thirds the average wage for males for more than two decades.[1] Numerous studies have attempted to explain this differential on the basis of differences in the male-female endowments of personal characteristics related to productivity (see, for example, Malcolm Cohen; Victor Fuchs; James Gwartney) Though these studies employ the rather extensive catalogue of personal characteristics measures available in modern micro-economic data files, none has explained even as much as one-half of the gross wage differential between the sexes Discrimination by employers against females has been suggested as the probable source of the unexplained residual, which amounts to roughly 20 percent of the average female wage rate (see, for example, Oaxaca, p 708).

That employer discrimination could account for such a sizeable depression in female wage rates for such an extended period of time is troublesome to anyone with even limited belief in the efficacy of the price mechanism If the services of female employees can indeed be purchased at a substantial discount below those of comparable male employees, why has there failed to appear an active group of firms that exploit this potential cost saving until only productivity related differentials remain?[2]

In the event that employers do not discriminate against women, there are two principal sources that may account for the unexplained male-female wage differential. The most apparent is the possibility that many important measures of male-female productivity differentials have not been included in existing empirical studies of earnings relationships. As additional data become available, it will be possible to investigate whether the unexplained wage differential shrinks in response to refinement and expansion of those personal characteristic measures that differ systematically between men and women.[3]

Alternatively, there may exist supply side phenomena that could account for equally qualified males and females being paid different wages by a nondiscriminating employer This paper focuses on the latter possibility and describes a very simple supply mechanism whereby nondiscriminating employers are expected to pay, on average, lower wages to females than to equally qualified males.

The point of departure is the observation that most adult participants in the labor force are married[4] and that husbands as a group work more hours and possess larger stocks of human capital– education, training, experience, and the like than do wives

For couples, the search for a pair of jobs

*Assistant professor of economics, Cornell University The first draft of this paper was prepared during my stay as a visiting fellow at the International Institute of Management in West Berlin I gratefully acknowledge the Institute's financial support, while exempting it from any responsibility concerning the study's contents.

[1] Ronald Oaxaca found, for example, that urban white males earned $2 95 per hour in 1967, as compared to $1.92 for urban white females

[2] Such a result would obtain even if most firms were willing to sustain income losses in order to discriminate against women, as long as nondiscriminating firms were numerous enough to be forced to compete with one another Fuchs has shown that the unexplained male-female wage differential is equally large for the self-employed as for other categories, a finding that is difficult to reconcile with the hypothesis that employer discrimination accounts for the gap

[3] Solomon W Polachek found, for example, that the greater irregularity of female labor force participation rates accounted for a significant portion of observed male-female wage differentials

[4] In March 1975, 71 percent of all men and 58 percent of all women in the labor force were married and living with their spouses

is constrained geographically: jobs ultimately accepted must both be in the same labor market.[5] In the course of this search, each spouse typically will generate a range of wage offers in several different labor markets. Only by coincidence will the best offers for both spouses occur in the same location. When they do not, one or both partners must compromise and accept something less than the best offer he or she has managed to generate.

Compromises of this sort will generally be more costly (in terms of family income and consumption) for the spouse who works more hours or whose stock of human capital is largest: on average, the husband. If the family's objective is to choose the pair of jobs in a single location that maximizes joint family income, it is then expected that husbands on average will make smaller compromises than their wives.[6] The differential degree of compromise that emerges from the requirement of family income maximization will in turn generate a pattern of wage differentials that is qualitatively the same as the one observed in existing wage studies: females will be observed to earn less, on average, than do males endowed with identical personal characteristics.[7]

A new model of the placement process is developed in Section I for the specific purpose of quantifying the family decision mechanism described above. Section II describes a procedure that can be employed to estimate the fraction of the previously unexplained wage gap that results from this family decision mechanism. An illustrative application of this procedure is presented in Section III. Section IV concludes with a list of remarks and qualifications concerning the interpretation of the model.

I. A Model of the Placement Process

A. *Characterizing Vacancies and Searchers*

The placement process is viewed as one in which employers and job seekers attempt to match the requirements of job vacancies with the personal characteristics of searchers. The requirements of a given job vacancy may be described in terms of the values taken by elements in a list of individual characteristics such as education, experience, intelligence, personality, and so on. In general, a vacancy's requirements do not specify fixed values for each characteristic, rather, a deficiency in one characteristic may be compensated for by an excess in another. In weighing experience against formal education, for example, employers may regard a vacancy as equally well-filled by any candidate possessing combinations of these attributes that lie along a particular locus, as depicted in Figure 1.

In Figure 1, any candidate whose com-

FIGURE 1. REQUIREMENTS OF A VACANCY WITH TWO CHARACTERISTICS

[5] S. Y. A. Ngai investigated the phenomenon of long distance commuting as a response to the geographical location constraint facing couples. Though growing, this phenomenon is currently of negligible importance.

[6] Single searchers do not operate under this type of geographic constraint and searchers of both sexes are thus free to locate in the market in which their best offer occurs.

[7] The idea that relative geographic immobility handicaps the labor market experience for women is not a new one. Beth Niemi, for example, argued that such immobility is one reason why women have higher unemployment rates than men. Nancy Gordon and Thomas Morton suggested that locational constraints result in a more inelastic supply curve for women than for men. In a "company town" setting, the monopsonist employer would then pay women lower wages than comparable qualified men. A differential of this sort could not arise, however, if the buyers' side of the labor market were competitive.

bination of experience and education places him along the vacancy description locus JJ is said to satisfy exactly the requirements of that particular vacancy, and employers are thus indifferent between different candidates on the JJ locus. Candidates to the southwest of JJ are deficient in their combination of education and training, while those to the northeast are overqualified with respect to these characteristics. In order to keep the notation compact, it is assumed that there is a single-valued correspondence between each vacancy description locus in the space of relevant characteristics and a scalar characteristics index, which will be denoted as J. Using the same indexing correspondence, each searcher is assigned a summary value S computed from his own particular values for the same list of characteristics used to describe vacancies In terms of these index values J and S, a searcher is exactly qualified for a particular vacancy when $J = S$, underqualified when $S < J$ and overqualified when $S > J$.

If no applicants are available for whom $S = J$ (as, in a probabilistic sense, there would almost never be), the employer is forced to select either an underqualified or overqualified applicant. In the former case, productivity is lower than in the case of an exactly qualified worker, and the employer could in principle compensate for this fact by offering underqualified applicants a lower wage rate. The excess qualifications of the overqualified applicant are by assumption of no productive value to the employer[8] and such applicants are offered the same wage that an exactly qualified applicant would receive

The exposition of the analysis that follows is considerably simplified (but none of the important conclusions are altered) if it may be assumed that employers do not hire underqualified applicants at all, restricting their consideration to only those applicants who are either exactly qualified or overqualified.[9] Also, it is convenient to choose the units in which the index J is measured in such a way that the wage offered to qualified candidates is related in some simple fashion to the value of J. Anticipating the form of the empirical wage relation of the next section, the units of J are chosen so that the wage corresponding to vacancy i is given by[10]

$$W_i = e^{J_i} \tag{1}$$

A job whose requirements are exceeded by the searcher will be accepted only if there is no better alternative, for acceptance of such a job implies a wage loss in comparison with the wage that could have been earned in a job for which he was exactly qualified. If X_i denotes the difference between an overqualified worker's qualification index S_i and the requirement index J_i of the job he accepts, using equation (1), his wage may be written as

$$W_i = e^{S_i}e^{-X_i} \tag{2}$$

Here the overqualified worker's wage is seen as the product of his wage in a job for which he is ideally suited and a factor less than unity representing his loss due to overqualification Equation (2) thus exhibits the hypothesized property that a given measure of overqualification is more costly (in terms of lost earnings) the higher is an individual's personal characteristics index S_i.

B. *The Search Process for Single Individuals*

The process of search for an individual may be viewed as identifying that vacancy in the market system for which he is least overqualified. Schematically, each labor market may be thought of as having a dis-

[8] There is some indication that individuals whose qualifications greatly exceed job requirements may actually be less productive than those who are slightly overqualified

[9] In defense of this assumption, very few among us would be willing to admit that our jobs make the maximum possible use of our talents; moreover, it may be said that this assumption provides a consistent description of the first $N - 1$ of the N jobs in a standard Peter Principle career sequence

[10] Gary Becker (1964) and Jacob Mincer have employed earnings relationships of this form in human capital studies

FIGURE 2. VACANCIES IN A SINGLE MARKET

tribution of vacancies with different requirements indices, as depicted below by the ordered collection of hash marks in Figure 2. The dotted hash mark in Figure 2 denotes the characteristics index of a particular searcher who, if search were limited to that single labor market, would maximize his wage by selecting vacancy J_i^*, the one for which the value X_i^* indicates that he is least overqualified. When search takes place in a number of markets at once, the individual simply selects that market in which J_i^* is largest (and in which the corresponding overqualification measure X_i^* is smallest).

In the present context, we wish to investigate the properties of the search process under equilibrium or near equilibrium conditions: that is, under conditions in which there is a reasonably close correspondence in both the number and in the range of characteristics values taken by both searchers and vacancies. Absent any concrete knowledge concerning the actual structure of vacancy and searcher characteristics for particular markets, the desired equilibrium conditions may be represented by viewing the collections of J and S values as random samples of equal size drawn from a common probability distribution.

For the single individual searching within this framework, the degree of overqualification X^* in the best job is also a random variable. The sampling distribution of X^* depends on N, the total number of vacancies in all markets combined. The exact form of this distribution also depends on the form of the distribution from which the J_i's and S_i's are drawn but is in no way influenced by the distribution of total vacancies between individual markets. For almost any reasonable distribution of the J_i's, the expected value of X^* for the single searcher decreases as N increases, approaching 0 in the limit.[11] Not surprisingly, the rich variety of opportunities characteristic of a large system of markets results in a low expected degree of overqualification for the single searcher, and, consequently, in a higher expected wage rate.

C. *The Search Process for Married Couples*

For the case in which search is confined to a single labor market, the description of the decision rule for a searching married couple that seeks to maximize its joint income is essentially the same as for single searchers. Each spouse selects the job for which he or she is least overqualified, as in the single searcher case.

When couples search over several geographically distinct labor markets, their decision rule is still simple to describe, but the equilibrium levels of overqualification that emerge for each spouse behave somewhat differently than do those for single searchers. For purposes of illustration, the case of a couple in which both spouses wish to work the same number of hours is considered. For such a couple searching in the multimarket case, the income-maximizing pair of jobs is the pair that maximizes the sum of their wage rates.

D. *Differential Overqualification*

To see how the expected overqualification levels of both spouses behave, it is useful to employ equation (2) to characterize the best pair of jobs as

$$J_M^*, J_F^* = J_{Mi}^*, J_{Fi}^* \tag{3}$$

for which

$$\max_{i=1,\ldots,k} \left\{ e^{S_M} e^{-X_{Mi}^*} + e^{S_F} e^{-X_{Fi}^*} \right\} \text{ obtains,}$$

where S_M and S_F denote the characteristics

[11] The rationale for this result may be seen easily by noting that, in terms of Figure 2 above, the expected distance between S_i and the closest vacancy hash mark to the left of S_i shrinks to zero as the quantity of vacancy hash marks increases.

indices for the husband and wife, and K denotes the number of markets over which the couple searches.

In equation (3), the couple's maximization exercise is seen as that of choosing the market whose pair of overqualification indices $X^*_{M_i}$, $X^*_{F_i}$ serves to maximize the weighted sum of the "ideal matchup" wage rates of the husband and wife, e^{S_M} and e^{S_F}. The precise nature of the dependence of the expectations of the optimal overqualification indices X^*_M, X^*_F on S_M, S_F is exceedingly complex, even when the underlying probability distribution from which the J_i's are drawn is simple in structure. Some important general properties of these expectations may be seen quite easily, however.

First, the expected overqualification levels, $E(X^*_M)$ and $E(X^*_F)$, will each exceed the expected value of X^* for a single searcher. It is expected, in other words that the income-maximizing pair of jobs will involve some compromise for both spouses, since only by chance will the best job for both spouses happen to occur in the same labor market.

Second, $E(X^*_M) < E(X^*_F)$ whenever $S_M > S_F$: the degree of overqualification expected for each spouse is ordered inversely to the ordering of their characteristics indices. This result follows from equation (2), which says that the cost of overqualification (in terms of foregone earnings) increases with the value of S. To the extent that husbands as a group have higher S values than wives, this differential in overqualification levels leads in turn to the outcome that a randomly chosen male will have higher expected earnings than a randomly chosen female with an identical personal characteristics measure. This result is reinforced when equation (3) is generalized to allow for the fact that husbands on average work more hours than wives.

Third, both $E(X^*_M)$ and $E(X^*_F)$ approach zero as the total number of vacancies in the market system approaches infinity. With enough markets and/or large enough markets to choose from, overqualification eventually ceases to be a quantitatively important phenomenon.

Fourth, the expected overqualification levels are not, as in the single-worker case, independent of the distribution of vacancies across individual markets. For a given level of total vacancies in the market system, $E(X^*_M)$ and $E(X^*_F)$ decrease as vacancies are more unevenly distributed across markets.[12] In the extreme case for which all vacancies are concentrated in a single labor market, the geographical search constraint is effectively eliminated and $E(X^*_M)$, $E(X^*_F)$ attain their minimum values for the given vacancy total.

E. *An Alternative Family Decision Rule*

Many may find the family income-maximization decision rule discussed above an implausible description of the way in which families decide where to live. An alternative family decision rule is one in which the wife accompanies her husband, locating in the geographic labor market that minimizes his overqualification level. For couples for whom S_M substantially exceeds S_F, or for whom the husband works substantially longer hours than his wife, this essentially male chauvinist decision rule produces expected male and female overqualification levels that are close to the expected overqualification levels associated with the family income-maximization rule.[13]

Whatever its rationale, the male chauvinist family location decision rule carries with it a number of analytically convenient implications about the equilibrium structure of expected male-female overqualification levels. For example, the expected over-

[12] This result relates directly to the observation that two-career couples have great difficulty securing satisfactory joint placement opportunities everywhere except in large urban areas. See, for example, the author

[13] For couples for whom both spouses' hours worked and S_M and S_F are close in value, and for whom the male chauvinist rule may conflict appreciably with family income-maximization considerations, the chauvinist rule may in many instances be assumed to have prevailed on the basis of historical notions about the relative psychosociological importance of career advancement for men as opposed to women

TABLE 1—MARKET SIZE AND EXPECTED OVERQUALIFICATION LEVELS UNDER THE MALE CHAUVINIST FAMILY LOCATION RULE

	Large Market	Small Market
Husbands	$E(X^*_{M1}) = \bar{X}^*_M(N)$	$E(X^*_{M0}) = \bar{X}^*_M(N)$
Wives	$E(X^*_{F1}) = \bar{X}^*_F(n_1)$	$E(X^*_{F0}) = (n_1/n_0)\bar{X}^*_F(n_1)$

Note: N = size of total market system; n_0 = size of small market; n_1 = size of large market.

qualification level of husbands becomes the same as for single workers. It declines with the total number of vacancies in the market system and is unrelated to the distribution of vacancies across individual markets. In particular, a husband observed working in a large labor market has the same expected overqualification level as a husband observed working in a small labor market.[14]

For wives operating under the male chauvinist location decision rule, no intermarket overqualification-minimization procedure takes place; such wives are in the position of having to do the best they can in the particular labor markets that have minimized their husbands' overqualification levels. If n_0 denotes the number of vacancies in the labor market chosen by a husband, his wife's overqualification level is given by

$$X_F^* = \min_{j=1,\ldots,n_0} X_{Fj} = X_F^*(n_0) \tag{4}$$

The expected overqualification level of the wife depends in a very simple way on the number of vacancies in the labor market in which her husband has located. If $\bar{X}_F^*(n_1)$ denotes the expected overqualification level in a market with n_1 vacancies, then $E(S_F^*)$ in a market with n_0 vacancies will be given by

$$\bar{X}_F^*(n_0) = \left(\frac{n_1}{n_0}\right) \bar{X}_F^*(n_1) \tag{5}$$

For example, when the number of vacancies doubles, the wife's expected overqualification level is cut by half, as illustrated in Figure 3.

As will become clear in the following section, the relationships between market size and expected overqualification levels play a pivotal role in the process of identifying the degree to which locational considerations dictate overqualification differentials and consequent wage differentials between the sexes. For the reader's convenience, these relationships are summarized in Table 1.

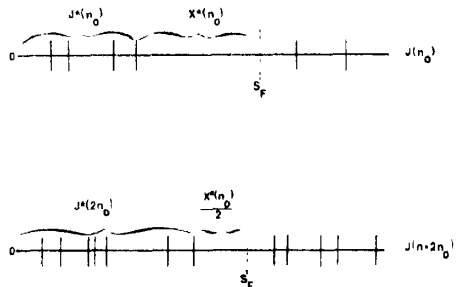



FIGURE 3. OVERQUALIFICATION LEVELS IN MARKETS OF UNEQUAL SIZE

II. Empirical Estimation of Overqualification Differentials

A. *Alternative Strategies*

The theory of overqualification differentials developed in the preceding section suggests a number of alternative procedures whereby one might attempt to estimate the portion of the unexplained male-female wage differential that arises because of family locational considerations.

Perhaps the most obvious of these involves a comparison of married and single

[14] The same observation applies equally well to single females and single males. While it is more probable *ex ante* that a single searcher's best placement will occur in a large market, there is no presumption *ex post* that observed large market placements of single workers will be superior in any way to small market placements.

females. Both the income-maximization and male chauvinist family decision rules suggest that single females, other things equal, should earn higher wages than married females, since the former are relatively more free from locational constraints. Considerable empirical evidence shows that single women do, in fact, tend to earn higher hourly wages than their married counterparts (see, for example, Fuchs; Polachek; Burton Malkiel and Judith Malkiel). Unfortunately, however, married and single women undoubtedly differ in many other important respects which are difficult to measure and which will inevitably blur any differential overqualification measure one might try to infer from a comparison of these two groups.

In order to avoid some of the problems inherent in comparisons of heterogeneous groups, an alternative estimating strategy may be considered in which the importance of overqualification differentials is inferred from the pattern of earnings differentials between married females working in labor markets of different size. For purposes of simplicity of notation and exposition, this strategy will be developed below for the case of two labor markets of unequal size, though it should be apparent that the two-market case is readily extended to the case of many markets.

B. *Heterogeneous Labor Markets, Male-Female Overqualification Differentials, and Sex Discrimination*

The case is considered in which two labor markets differ (for the purpose of wage determination) in both size and in some other unknown respect. The assumption is made that married females have located in these markets as a result of their husbands having found best jobs there. As noted in the previous section, this assumption is broadly consistent with the requirements of family income maximization for possibly a large majority of families, and implies that the expected overqualification levels for wives will vary inversely with the size of the market in which they have located.

In the discussion that follows, a_0 and a_1 are used to denote market-specific effects on wage rates of earners in the small and large markets, respectively. These effects are distinct from overqualification effects and for concreteness may be thought of as representing a compensating differential attributable to some environmental difference between the two locations.

Using the notation developed in the previous section, the logarithms of the wage rates received by married earners in the two markets may then be written as

$$\ln W_{F0_i} = a_0 + S_{F0_i} - X^*_{F0_i} \quad (6)$$

$$\ln W_{M0_i} = a_0 + S_{M0_i} - X^*_{M_i} \quad (7)$$

$$\ln W_{F1_j} = a_1 + S_{F1_j} - X^*_{F1_j} \quad (8)$$

$$\ln W_{M1_j} = a_1 + S_{M1_j} - X^*_{M_j} \quad (9)$$

for spouses in the ith and jth couples in the small and large markets, respectively. In equation (6), for example, W_{F0_i} denotes the hourly wage rate for the wife in the ith couple located in the small market, S_{F0_i} is her characteristics index and $X^*_{F0_i}$ is her overqualification level. Market subscripts are omitted from the husbands' overqualification levels, $X^*_{M_i}$, $X^*_{M_j}$, because these do not depend on market size.

The standard procedure in empirical wage investigations has been to approximate the true productivity measure as a weighted sum of such characteristics as age, schooling, experience, and so on. If Z denotes a vector of observable personal characteristics and β a vector of coefficients, the relationship between the true characteristics indices S_{F_i} and S_{M_i}, and their measured values may be represented by

$$S_{F0_i} = S_{F1_i} = f_i + Z_i'\beta + \epsilon_i \quad (10)$$

and

$$S_{M0_j} = S_{M1_j} = m_j + Z_j'\beta + \epsilon_j \quad (11)$$

where ϵ_i, ϵ_j are stochastic disturbance terms with mean zero, and where f_i and m_j represent the influence of unobserved personal characteristics for the ith female and jth male, respectively.

Substituting equations (10) and (11) into

equations (6)–(9) we generate the empirical wage relationship for the pooled sample of husbands and wives in the two labor markets:

$$(12)\quad \ln W_i = a_0(D_{F0i} + D_{M0i}) + a_1(D_{F1i} + D_{M1i}) + Z_i'\beta + [\epsilon_i + (D_{M0i} + D_{M1i})(m_i - X_M^*) + D_{F0}(f_i - X_{F0i}^*) + D_{F1}(f_i - X_{F1i}^*)] = a_0(D_{F0i} + D_{M0i}) + a_1(D_{F1i} + D_{M1i}) + Z_i'\beta + \gamma_i,$$

where

D_{F0i} = 1 for small market wives, 0 otherwise
D_{M0i} = 1 for small market husbands, 0 otherwise
D_{F1i} = 1 for large market wives, 0 otherwise
D_{M1i} = 1 for large market husbands, 0 otherwise

The random disturbance term γ_i in the wage relation (12) has a nonzero mean given by

$$(13)\quad \bar{\gamma} = (D_{M0i} + D_{M1i})(\bar{m} - \bar{X}_M^*) + D_{F0i}(\bar{f} - \bar{X}_{F0}^*) + D_{F1i}(\bar{f} - \bar{X}_{F1}^*)$$

where $\bar{m}$ and $\bar{f}$ are the means of the unobserved characteristics values for husbands and wives, respectively. It is convenient to rewrite the wage relation (12) as

$$(14)\quad \ln W_i = a_{F0}D_{F0i} + a_{M0}D_{M0i} + a_{F1}D_{F1i} + a_{M1}D_{M1i} + Z_i'\beta + (\gamma_i - \bar{\gamma})$$

where the coefficients of the location-sex dummies are defined as

$$(15)\quad a_{F0} = a_0 + \bar{f} - \bar{X}_{F0}^*$$

$$(16)\quad a_{M0} = a_0 + \bar{m} - \bar{X}_M^*$$

$$(17)\quad a_{F1} = a_1 + \bar{f} - \bar{X}_{F1}^*$$

$$(18)\quad a_{M1} = a_1 + \bar{m} - \bar{X}_M^*$$

The disturbance term in the rearranged wage equation (14) has the zero mean property assumed in conventional estimating procedures. If overqualification levels are not systematically related to individual personal characteristics and if the original disturbance terms ϵ_i obey the standard assumptions, then ordinary least squares (*OLS*) will yield unbiased, consistent estimates of the four sex-location intercepts in equation (14).[15]

The estimates of these intercepts can in turn be used to construct estimates of the relative degree to which overqualification depresses wives' wage rates in the two markets. Except for the fact that wage rates are assumed to differ between the two markets, not only because of overqualification differentials but for other reasons as well, the effect of overqualification on female wages could be inferred from a direct comparison of the intercepts for wives in the two markets. Under the present circumstances, however, we must first eliminate local effects by comparing the differentials between the male and female intercepts for the large and small markets:

$$(19)\quad (a_{M0} - a_{F0}) - (a_{M1} - a_{F1}) = \bar{X}_{F0}^* - \bar{X}_{F1}^*$$

This difference measures the difference in wives' wages in the two markets that is attributable to the fact that wives in the smaller market are expected to have higher overqualification levels than are large-market wives. Using the property (from equation (5) above) that average female overqualification levels in the two markets are inversely proportional to the size of the two markets, we have

$$(20)\quad (a_{M0} - a_{F0}) - (a_{M1} - a_{F1}) = \bar{X}_{F0}^*(1 - n_0/n_1)$$

If $\hat{a}_{M0}, \hat{a}_{F0}, \hat{a}_{M1}, \hat{a}_{F1}$ denote the *OLS* estimators of the location-sex intercepts, then

$$(21)\quad \hat{\bar{X}}_{F0}^* = \frac{n_1}{n_1 - n_0}\{(\hat{a}_{M0} - \hat{a}_{F0}) - (\hat{a}_{M1} - \hat{a}_{F1})\}$$

[15]Because the pooled regression has heteroscedastic disturbances, generalized least squares (*GLS*) will afford more efficient estimators than *OLS*. In view of the prodigious size of recent micro data banks, however, this consideration is of little practical significance.

is a consistent, unbiased estimator of the degree to which wives' wages in the small market are depressed by overqualification.

Given the logarithmic form of the wage relation (14), the estimate of $\bar{X}^*_{F0}$ given in equation (21) may be interpreted as the percentage by which wages are reduced by overqualification in the small market for wives whose overqualification level equals the average level for that market.

Under the assumption that the location of the husband's best job dictates the family location decision, the expected level of overqualification for husbands in the small market is the same as it would be in any other market. The $\bar{X}^*_M$ depends only on the total number N of vacancies in the market system as a whole, and is related to $\bar{X}^*_{F0}$ by

$$(22) \qquad \bar{X}^*_M = \frac{n_0}{N}\bar{X}^*_{F0}$$

The difference

$$(23) \qquad \bar{X}^*_{F0} - \bar{X}^*_M = \bar{X}^*_{F0}(N - n_0)/N$$

represents the expected overqualification differential between husbands and wives in the small market and expresses the cost, as a percentage of the average wage rate, of the wife having constrained her search of her husband's best job site. An estimate of this overqualification differential can be constructed from the previous estimate of $\bar{X}^*_{F0}$:

$$(24) \qquad (\hat{X}^*_{F0} - \bar{X}^*_M) = ((N - n_0)/N)\hat{\bar{X}}^*_{F0}$$

The size of the total market system over which couples search will in practice vary from couple to couple, but for most will be considerably larger than the two individual markets considered above. If N is indeed large in relation to n_0, then the expected overqualification level for the husband becomes insignificant in relation to that of his wife, and the overqualification differential for the average couple in the small market is reasonably approximated by the wife's average overqualification estimate given in equation (21).

In order to determine the fraction of the husband-wife wage gap that is explained by overqualification differentials, it remains to compute the fraction of that gap attributable to unobserved differences in husband-wife personal characteristics or to employer discrimination against wives. An expression for the latter magnitude, which is defined by the difference

$$(25) \qquad r = \bar{m} - \bar{f}$$

can be constructed from the husband and wife intercepts for either market. From the difference $(a_{M0} - a_{F0})$ in equations (15) and (16) we first solve for

$$(26) \qquad r = a_{M0} - a_{F0} + (\bar{X}^*_{F0} - \bar{X}^*_M)$$

If it may again be assumed that husbands' overqualification levels are on average insignificant in relation to those of wives in the small market, a suitable estimate of r is given by

$$(27) \qquad \hat{r} = \frac{n_1}{n_1 - n_0}(\hat{a}_{M1} - \hat{a}_{F1}) - \frac{n_0}{n_1 - n_0}(\hat{a}_{M0} - \hat{a}_{F0})$$

III. An Illustrative Application

A. *The Sample*

This section describes the results of applying the procedures developed in the preceding section to a sample of urban husbands and wives selected from the 1967 Survey of Economic Opportunity (*SEO*).

The phenomenon that produces male-female overqualification differentials applies more to certain population groups than to others. In less skilled occupations, for example, vacancy requirements are much less likely to exhibit the kind of dispersion that would result in significant overqualification differentials than are those in more highly skilled occupations. For this reason, the sample was limited to those individuals in the "professional, technical, and kindred" and "managers, officials, and proprietors" occupational categories.

Because overqualification differentials are the outcome of geographical mobility, such differentials will not be observed for couples

for whom economic considerations play no role in the location decision. In the hope of limiting the number of such couples in the sample, no couple was included whose current residential location was the same as the husband's location at age 17. The sample was further restricted to include only nonrural private wage and salary workers for whom health considerations restricted neither the kind nor amount of work performed. These restrictions resulted in a sample of 444 husbands and 89 wives.

TABLE 2—URBAN SIZE CATEGORIES

Urban Category	Average 1970 Population (1000's)
Urban Non-*SMSA*	12.0
SMSA < 250,000	144.6
SMSA 250,000 – 500,000	329.3
SMSA 500,000 – 750,000	620.8
SMSA 750,000 +	2138.9

Source. U.S. Bureau of the Census, *Census of Population: 1970*, v. 1, Part A

B. *The Wage Equation*

The *SEO* data permit the identification of five urban size categories for each respondent as shown in Table 2. Separate intercepts for husbands and wives were estimated for each of these five urban size categories. Control variables were also included in the wage equation and are shown in Table 3. Separate experience coefficients were estimated for mothers in an attempt to capture the effect on earnings of the irregular labor force participation pattern associated with the early years of childrearing (see, for example, Polachek). The results of fitting the above wage equation to the sample of highly skilled urban married persons are reproduced in Table 4.

The pattern of wives' wages increasing with market size predicted by the overqualification differentials hypothesis[16] is confirmed qualitatively by the urban size coefficients reported in Table 4. The pattern of husbands' urban size coefficients shows no appreciable variation across the first three market categories but, contrary to prediction, the two largest categories show significantly higher wages than for urban non-Standard Metropolitan Statistical Areas (*SMSA*).[17]

The fact that separate intercepts were estimated for each of several market size categories means that the overqualification differential between husbands and wives can be estimated in a variety of ways. Designating *SMSA* < 250,000 as the reference category (hereafter, small *SMSA*), four separate estimates of $\bar{X}^*_{f0}$ (the average overqualification level for wives in small *SMSA*s) can be computed by comparing the difference in the husbands' and wives' intercept terms for the reference market with the corresponding differences for each of the other four markets, according to equation (21).[18] These four estimates were computed and found to exhibit a rather disturbing degree of dispersion, ranging from a high of more than .23 to a low of less than .01. Accordingly, any inferences about overqualification differentials based on the coefficients reported in Table 4 should be considered highly tentative.

[16]Because each urban category variable includes a large number of physically distinct markets, there is a presumption that market-specific effects that are unrelated to market size will not be distinguishable in the values of the urban category coefficients. Differences in wives' wage rates across categories will thus be the result either of overqualification differentials or of some other cause that is specifically related to market size.

[17]Such an increase could, for example, be the result of higher living costs characteristic of larger cities. Because overqualification differentials are inferred from intermarket comparisons of intramarket husband-wife earnings differences, however, cost-of-living differentials should introduce no distortions. Alternatively, the higher husbands' wages observed in large cities could result from some unmeasured differences in the personal characteristics of this group. Differences of this sort would bias downward our estimates of overqualification differentials.

[18]The assumption is made that the number of job vacancies in each market is proportional to the market's population.

TABLE 3 —CONTROL VARIABLES

Symbol	Definition	Mean Value
EDUC	Years of school completed	14.52
*EDUC*²		218.68
EXPER	Age-Educ-6 for husbands and for wives without children[a]	10.57
*EXPER*²		153.53
MEXPER	Age-Educ-6 for wives with children	1.77
*MEXPER*²		28.12
Census Region	Northeast, Northcentral, South, or West	

[a]The *SEO* file unfortunately contains no real information on individual work experience. Oaxaca has employed the difference between (age) and (the number of years of education + 6) as a measure of the maximum length of work experience attainable by an individual of given age. "Age-Educ-6" is used for the same purpose here

TABLE 4 DETERMINANTS OF WAGE RATES FOR HIGHLY SKILLED URBAN MARRIED PERSONS

<table>
<tr><th></th><th>Male</th><th>Female</th></tr>
<tr><td>Urban non-SMSA</td><td>000</td><td>– 355</td></tr>
<tr><td></td><td></td><td>(.154)</td></tr>
<tr><td>SMSA < 250,000</td><td>– 014</td><td>– 347</td></tr>
<tr><td></td><td>(082)</td><td>(178)</td></tr>
<tr><td>SMSA 250,000 – 500,000</td><td>– 011</td><td>–.214</td></tr>
<tr><td></td><td>(087)</td><td>(170)</td></tr>
<tr><td>SMSA 500,000 750,000</td><td>197</td><td>– 128</td></tr>
<tr><td></td><td>(088)</td><td>(170)</td></tr>
<tr><td>SMSA, 750,000 +</td><td>185</td><td>– 094</td></tr>
<tr><td></td><td>(069)</td><td>(116)</td></tr>
<tr><td>EDUC</td><td colspan="2">230</td></tr>
<tr><td></td><td colspan="2">(059)</td></tr>
<tr><td>EDUC²</td><td colspan="2">– 006</td></tr>
<tr><td></td><td colspan="2">(002)</td></tr>
<tr><td>EXPER</td><td colspan="2">035</td></tr>
<tr><td></td><td colspan="2">(015)</td></tr>
<tr><td>EXPER²</td><td colspan="2">– 00039</td></tr>
<tr><td></td><td colspan="2">(00061)</td></tr>
<tr><td>MEXPER</td><td colspan="2">.057</td></tr>
<tr><td></td><td colspan="2">(022)</td></tr>
<tr><td>MEXPER²</td><td colspan="2">– 0019</td></tr>
<tr><td></td><td colspan="2">(00097)</td></tr>
<tr><td>Northeast</td><td colspan="2">000</td></tr>
<tr><td>Northcentral</td><td colspan="2">– 080</td></tr>
<tr><td></td><td colspan="2">(055)</td></tr>
<tr><td>South</td><td colspan="2">- 126</td></tr>
<tr><td></td><td colspan="2">(056)</td></tr>
<tr><td>West</td><td colspan="2">–.103</td></tr>
<tr><td></td><td colspan="2">(.057)</td></tr>
<tr><td>Constant</td><td colspan="2">1.130</td></tr>
</table>

$R^2 = .27$
Sum of squared residuals = 34.68
Number of observations = 533

Note: Standard errors are in parentheses Dependent variable: logarithm of hourly wage rate

With this caveat in mind, the arithmetic average of these four estimates was computed, yielding the following estimate for the average overqualification level of wives in small *SMSA* (which, in 1970, had an average population of roughly 145,000):

$$(28) \qquad \overline{X}^*_{F0} = .078$$

Taken at face value, this estimate says that wives in skilled occupations in small SMSAs earn nearly 8 percent less than husbands with the same measured levels of education and experience. This amounts to nearly a quarter of the small *SMSA* husband-wife earnings differential that is unexplained by the crude and limited set of personal characteristics control variables in Table 4. For all of its uncertainty, this figure appears substantial enough to suggest that the subject of overqualification differentials is worthy of more detailed empirical exploration.

IV. Conclusions

The subject of male-female earnings differentials is one charged with considerable emotion and controversy. Accordingly, it is important to spell out very clearly just what the magnitudes described in the previous sections do and do not imply.

The average overqualification differential between spouses in a labor market of representative size, $\overline{X}^*_{F0} - \overline{X}^*_M$, measures the average fraction by which wives' earnings

are reduced purely as a result of having followed their husbands to a particular geographic location. This component of the observed male-female earnings differential is not the result of discrimination against females *by employers* and in general will be unresponsive to attempts to eliminate employer discrimination.[19] To a large extent, however, differences in the male-female distributions of productivity related characteristics make outwardly sexist family decision rules economically rational, and one must look beyond the family in examining the extent to which antifemale bias accounts for the uneven distribution of productive characteristics endowments we presently observe.

The expression for the residual parameter r given in equation (27) may be described as an estimate of the gap between male and female wages (expressed as a fraction of the overall average wage rate) that cannot be accounted for on the basis of differences in observable personal characteristics or by family location considerations of the type discussed in this study. To the extent that employers do discriminate against females, effects of such behavior on wage differentials will be included in this residual. The residual will also include the effect of differences in unobservable personal characteristics between males and females.

Personal characteristics that are unobservable by researchers may or may not be observable by employers. Employers may well be aware, for example, of how much firm-specific human capital an employee has, whereas researchers are largely ignorant of this magnitude. Wage differentials that arise from male-female differences in the values of such characteristics will be similar to those that arise from differences in observables such as formal education, in the sense that neither constitutes a signal of antifemale discrimination by employers. The influence of unobservable characteristics, however, will be visible to researchers only as part of the residual estimate, where it will be empirically indistinguishable from the influence of discrimination.

It is also possible that characteristic differentials that are observable for individuals by neither employers nor researchers may be responsible for some of the otherwise unexplained male-female wage differential. If, for the sake of argument, employers have observed from experience that males are more productive than females with similar observable personal characteristics, then females will fall victim to the type of statistical discrimination described by Edmund Phelps: a female with true productivity characteristics identical to those of a given male will be paid less even by cost-minimizing employers if all relevant characteristics are not observable but must instead be estimated on the basis of male and female averages inferred from historical experience. As in the case of the effects of characteristics observable by the employer but not the researcher, such statistical discrimination effects will show up as part of the unexplained residual estimate given in equation (27), where they too are inseparable from the effects of pure sex discrimination.

While all of the factors that unite to form the unexplained residual term are separate and distinct from the influence of overqualification differentials, some of these factors may themselves be influenced by the family location decision mechanism that generates these differentials. In the human capital literature, for example, firms are described as willing to invest in firm-specific human capital for an employee under the assumption that he will not leave the firm before its investment is recovered. For males operating under the family location mechanism described in this study, this is likely to be a reasonable assumption, since, by definition, firm-specific human capital is worth less to the employee in firms other than his own. For females under the same

[19]This, of course, does not imply that the particular type of overqualification differentials discussed in this study are not the result of discrimination against females by persons or institutions other than employers. Indeed, insofar as sexist family location decision mechanisms are alone responsible for these overqualification differentials, these are very palpably the result of antifemale bias

decision mechanism, however, this calculus clearly may not apply: if wives do indeed move or stay less in response to considerations involving their own careers than to considerations involving their husbands', cost-minimizing employers will respond by investing in less firm-specific human capital for women.[20]

The identification and disentanglement of these and other possible effects in the estimate of the unexplained residual in the male-female differential will have to await the development of more comprehensive data and better methods of analysis, but at least a much clearer picture of their combined total size should be made possible by eliminating the effects of locationally determined overqualification differentials

[20] By this line of reasoning, the incentives for firms to invest in general human capital should be equally low for both men and women If the acquisition of a home and family makes married males less geographically mobile than single males, firms would be more willing to invest in specific traning for married males, which might be one reason why married males earn more on average than single males

REFERENCES

E. K. Allison, "Sex Linked Earnings Differentials in the Beauty Industry," *J Hum Resources*, Summer 1976, *11*, 383 90.

K. Arrow, "Models of Job Discrimination," in Anthony H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972.

Gary S. Becker, *Human Capital: A Theoretical and Empirical Analysis*, New York 1964.

———, *The Economics of Discrimination*, 2d ed., Chicago 1971

G. D. Brown, "How Type of Employment Effects Earnings Differences by Sex," *Mon. Labor Rev.*, July 1976, *99*, 25-30.

J. E. Buckley, "Pay Differences Between Men and Women in the Same Job," *Mon. Labor Rev.*, Nov. 1971, *94*, 36–40.

M. S. Cohen, "Sex Differences in Compensation," *J. Hum. Resources*, Fall 1971, *6*, 434–47.

R. H. Frank, "Family Location Constraints and the Geographic Distribution of Female Professionals," *J. Polit. Econ.*, Feb. 1978, *86*, 117–30.

V. R. Fuchs, "Differences in Hourly Earnings Between Men and Women," *Mon. Labor Rev.*, May 1971, *94*, 9–15.

H. N. Fullerton and J. J. Byrne, "Length of Working Life for Men and Women," *Mon Labor Rev.*, Feb. 1976, *99*, 31–35.

N. M. Gordon and T. E. Morton, "A Low Mobility Model of Wage Discrimination With Special Reference to Sex Differentials," *J. Econ. Theory*, Mar. 1974, *7*, 241 53.

J. Gwartney, "Discrimination and Income Differentials," *Amer. Econ. Rev.*, June 1970, *60*, 396–408.

T. Johnson, "Returns from Investment in Human Capital," *Amer. Econ. Rev.*, Sept. 1970, *60*, 546 60.

H. Kahne, "Economic Perspectives on the Roles of Women in the American Economy," *J. Econ. Lit.*, Dec. 1975, *13*, 1249 292.

Janice F. Madden, *The Economics of Sex Discrimination*, Lexington 1973.

B. G. Malkiel and J. A. Malkiel, "Male-Female Pay Differentials in Professional Employment," *Amer. Econ. Rev.*, Sept. 1973, *63*, 693–705.

J. Mincer, "The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach," *J. Econ. Lit*, Mar. 1970, *13*, 1 26.

——— **and S. W. Polachek,** "Family Investment in Human Capital: Earnings of Women," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, Part II, S76-S108.

S. Y. A. Ngai, "Long Distance Commuting as A Solution to Geographic Limitation to Career Choices of Two Career Families," masters thesis, Mass. Instit. Technology 1974.

B. Niemi, "Geographic Immobility and Labor Force Mobility: A Study of Female Unemployment," in Cynthia B. Lloyd, ed., *Sex Discrimination and the Division of Labor*, New York 1975.

R. Oaxaca, "Male-Female Wage Differentials in Urban Labor Markets," *Int. Econ. Rev.*, Oct. 1973, *14*, 693-709.

E. S. Phelps, "The Statistical Theory of Racism and Sexism," *Amer. Econ. Rev.*, Sept. 1972, *62*, 659-61.

S. W. Polachek, "Discontinuous Labor Force Participation and Its Effects on Women's Market Earnings," in Cynthia B. Lloyd, ed., *Sex Discrimination and the Division of Labor*, New York 1975.

H. Sanborn, "Pay Differences Between Men and Women," *Ind. Labor Relat. Rev.*, July 1964, *17*, 534-50.

S. P. Smith, "Government Wage Differentials by Sex," *J. Hum. Resources*, Spring 1976, *11*, 185-99.

J. E. Stiglitz, "Approaches to the Economics of Discrimination," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 287-95.

R. P. Strauss and F. W. Horvath, "Wage Rate Differences by Race and Sex in the U.S. Labor Market: 1960-1970," *Economica*, Aug. 1976, *43*, 287-98.

U.S. Bureau of the Census, *U.S. Census of Population: 1970, 1*, Part A.

U.S. Department of Health, Education, and Welfare, Office of Economic Opportunity, *Survey of Economic Opportunity*, Washington 1967.

Optimal Investment Strategy for Boomtowns: A Theoretical Analysis

By RONALD G. CUMMINGS AND WILLIAM D. SCHULZE*

As the United States continues its drive for self-sufficiency in energy, a wide range of associated problems have appeared such as public safety (for example, the disposition of uranium tailings), environmental quality, competition for resources between the energy sector and other sectors (particularly, the competition for water with the agricultural sector), and the creation of boomtowns. Current manifestations of these problems are most likely only the tip of the iceberg given the unprecedented scale of new energy developments anticipated in the United States, particularly in the western states (see Federation of Rocky Mountain States). Increasing concern has been raised as to just how state, federal, and local governments are to deal with these problems given the substantial shift in the basic structure of the *U.S.* energy producing sector.

Of these problems, those associated with boomtowns are the subject of growing concern for policymakers in Rocky Mountain states, reflecting to some extent the recent discouraging experiences witnessed in such communities as Rock Springs and Gillette, Wyoming (see John Gilmore and Mary Duff). Boomtowns are characterized by large jumps in population over a five- to twelve-year construction period, after which the population settles to a lower level associated with the completed activities such as mining or power plant operation. The strains placed on ofttimes fragile sociocultural and institutional systems, as well as on social infrastructure, which result from such rapid rates of population increase in relatively sparsely populated areas, are apparent after only a moment's reflection, and have been described elsewhere in some detail (see Gilmore, and Gilmore and Duff).

In this paper we are concerned with that dimension of boomtown problems related to the provision of social infrastructure (streets, roads, schools, public safety facilities, etc.) Given a community faced with boomtown conditions, it seems clear that the community would not wish to invest in infrastructure in an effort to maintain preboom period per capita levels for the relatively short-lived peak of the boom inasmuch as substantial idle stocks would remain in future years. On the other hand, it is not at all clear that per capita infrastructure is optimally held at the lower levels corresponding to the long-run (postconstruction) "stable" population. On intuitive grounds one would expect an *optimal* policy for investments in social infrastructure to be one wherein the marginal capital costs of excess capacities during the postconstruction period are in some sense equated with the marginal social costs attributable to deteriorated services from somewhat lower levels of social infrastructure during the construction phase.

In considering these problems, intuition is virtually the only game in town at the present. Looking to existing literature, only the most limited efforts seem to have been made to date in terms of providing a conceptual structure which might serve as an analytical framework for this class of problems (see Berry Ives, William Schulze,

*Professor of economics, University of New Mexico, and associate professor of economics, University of Southern California, respectively We wish to acknowledge the financial support provided for this work by the National Science Foundation through the RANN Lake Powell Research Project and to the Los Alamos Scientific Laboratory. We express our appreciation to Allen Kneese and Karl-Göran-Mäler for their helpful comments on an earlier draft All opinions and remaining errors are of course our sole responsibility.

and David Brookshire; Cummings and Arthur Mehr).

The central purpose of this paper is thus to suggest such a conceptual structure and to deduce from this structure a number of lines of argument which have clear operational significance in terms of future research. To this end, in Section I we construct a general investment model which, drawing on relevant aspects of welfare economics, captures the essence of the boomtown's decision environment. In Section II the economic characteristics of a socially optimal investment strategy are analyzed in some detail, and policy ramifications related to tax policies, construction scheduling (for the energy facility), and financing (for example, "front-end" monies) are discussed. Results from our optimization model are shown to provide a basis for explaining various characteristics of ongoing booms. Greater stability (in terms of periodic changes in population and the labor force) during the construction phase is shown to occur the lower are interest rates and the *sooner* that investments in infrastructure occur. Investments in social infrastructure that begin only after the construction phase is well under way are shown to potentially *exacerbate* the boom, increasing instability rather than reducing it.

Concluding remarks which focus on the ramifications of these results for empirical research are given Section III. The notation used in the sections below is as follows:

t = time
t^* = the time at which investments in social infrastructure cease
W = wage rate
L = labor used for construction of the facility, variable over the construction interval $0 \leq t \leq T$
$\bar{L}$ = labor used for operation of the facility, constant during the operating interval $T \leq t \leq \infty$
I = investment in social infrastructure
K = social capital (where per capita stocks are given as $k = K/L$)
$K^-(0)$ = the initial stocks of social infrastructure in the community before construction of the facility begins
$K^+(0)$ = the initial stock of social infrastructure after front-end investments
$f(L)$ = a concave production function for construction of the facility
$C(t)$ = construction completed at time t ($\bar{C}$ = completion)
r = social rate of discount
m = maintenance and operating costs/unit of K
$U(k, W)$ = utility of individuals (U is a concave function, $U_k, U_W > 0$)
h = the (constant) elasticity of labor in construction
η = elasticity of substitution between wages and per capita infrastructure.

I. An Investment Model for Boomtowns

We begin by positing the existence of a small municipality in or around which a large-scale energy extraction/conversion facility is to be located. Our decision maker is assumed to control not only investment policies for social infrastructure in the municipality, but the staging of construction activities at the new facility as well. It follows that two types of capital stocks and investments are of concern to the decision maker. One type of capital investment is social infrastructure, the other is the facility under construction.

The stock of social infrastructure at any time t is defined as $K(t)$ where additions to this stock at t are $I(t)$. We assume that, with a rate of periodic unit maintenance costs m, capital stocks are of infinite life. The intertemporal transition of capital stocks is then given by

$$(1) \quad dK(t)/dt = I(t) \geq 0, \quad 0 \leq t \leq T$$

Note that we are assuming that capital investment is irreversible in that we treat $I(t)$ as nonnegative. This assertion is reasonable for roads, schools, and public

buildings, and characterizes the central problem of boomtowns: the immobility of much of the appropriate social infrastructure.

We assume that the construction of the new facility is to be completed by time T. The level of the facility's completion at any t is given by $C(t)$, expressed, for example, in dollars. Periodic construction is determined by the amount of labor used $L(t)$, and is given by a concave function $f(L)$. Defining $\bar{C}$ as the completed facility, we then require the following:

(2)
$$dC(t)/dt = f(L), f' \geq 0, f'' \leq 0, 0 \leq t < T$$

(3)
$$C(T) = \bar{C}$$

In terms of acquiring the required labor for construction L, we assume a utility function for homogeneous units of labor of the form $U(k, W)$, where W is the wage and k is per capita (in units of labor)[1] social infrastructure, K/L. The inclusion of k in the utility function implies the use of k as a surrogate for the individual's "quality of life" in a community, and clearly a number of considerations other than social infrastructure are relevant for such quality. The ramifications of introducing an expanded set of variables relevant to the utility of labor are discussed below in Section III

We define $U(\bar{k}, \bar{W})$ as a representative utility level for labor in communities from which our boom community must attract its labor supply where $\bar{k}$ and $\bar{W}$ characterize social capital and wage levels in these communities. To induce labor to locate in our boom community, we must then require that

(4)
$$U(k, W) \geq U(\bar{k}, \bar{W}), U_k, U_W \geq 0$$

where we abstract from transportation costs and assume that equality holds in the first relation.

[1]With population P some function $P(L)$ of the labor force, per capita infrastructure is of course $K/P(L)$. The P may be viewed as linear in L, and we ignore this transformation of L to P to simplify the analysis.

From (4) it then follows that $dW/dk = -(U_k/U_W) < 0$, and a tradeoff exists between boomtown wages and per capita social capital. Of course, this tradeoff lies at the heart of the inquiry intended here and implies that wages become a function of per capita social capital stocks, $W(k)$.

With the structural conditions (1)-(4), we assume that our decision maker wishes to minimize the social costs associated with the construction and operation of the new facility, where social costs are taken to be the sum of the wage bill for constructing and operating the facility ($W(k)L(t)$), plus investment (I) and maintenance costs (mK) for social capital. Note then that maintenance costs are assumed to be a constant fraction m of capital costs. We also allow for the possibility that the decision maker may wish to preinvest in social capital to provide an initial stock ($K^+(0)$) greater than which currently exists ($K^-(0)$). For the postconstruction period $\infty \geq t \geq T$, we assume a constant labor force $\hat{L}$ which is required for the operation of the facility, in which case social costs during the postconstruction periods depend solely on social capital at T; that is, the present value of social costs during the period $\infty \geq t \geq T$ are given by

(5)
$$\int_T^\infty e^{-rt}[W(K(T)/\hat{L})\hat{L} + mK(T)]dt$$

Assuming as we do that equilibrium conditions apply in the interval $t \geq T$ with L fixed at $\hat{L}$, the wage W that results in equality in (4) is determined solely by $K(T)$, the remaining capital stock (where $W(k) = W(K(T)/\hat{L})$).

Our decision maker is thus seen to face the problem of choosing nonnegative values for $L(t)$ and $I(t)$ which minimize the following expression subject to the conditions (1)-(3).

(6)
$$\text{Min: } K^+(0) - K^-(0) + \int_0^T e^{-rt}[W(k)L(t) + I(t) + mK(t)]\,dt + \int_T^\infty e^{-rt}[W[K(T)/\hat{L}]\hat{L} + mK(T)]dt$$

The decision maker's motivation for minimizing (6) stems from his assumed control over capital stocks used for infrastructure and the facility. Increments to capital stocks for the facility obtain via increases in L(t) which by (4) imply higher wages, and therefore higher "investment" costs for the facility. Increments to infrastructure via I(t) directly imply higher costs for infrastructure, but lower wages (investment costs for the facility) will result (see (4)). Basic to the minimization of (6) is thus the wage-infrastructure tradeoff alluded to above.

Of course, the structure of (6) allows for investments in infrastructure *prior* to arrival of the construction labor force ($K^+(0)$). Further, the right-most integral (from T to infinity) is essentially a terminal value function measuring the infrastructure–related benefits that accrue to the postconstruction period labor force.

Necessary conditions for a minimum of (6) are developed in the Appendix. To facilitate the discussion in the next section, we wish to state a few key propositions, proofs for which are also given in the Appendix. These propositions make use of two additional assumptions that greatly simplify the analysis. They are:

Assumption 1: Constant elasticity of substitution between wages and per capita social infrastructure in equation (4):

$$(7) \qquad \eta = -\frac{dW}{dk}\frac{k}{W} \qquad \eta > 0$$

Assumption 2: Constant elasticity of labor in construction in equation (2):

$$(8) \qquad h = \frac{df}{dL}\frac{L}{f} \qquad 1 > h > 0$$

The term h is of course also the share of labor in the competitive construction industry.

With these additional assumptions, the following propositions hold:

PROPOSITION 1: *During any interval in which labor is used for construction and Assumption* 1 *holds, the marginal revenue product of labor in construction exceeds the construction wage by a factor of* $1 + \eta$. *Thus, we have*

$$(9) \qquad (1 + \eta)W = \bar{\phi}e^{rt}f'(L)$$

where $\bar{\phi}e^{rt}$ *is the current value of construction at time* t. (The symbol ϕ is used for the costate variable on (2) above in the optimal control problem as specified in the Appendix.)

PROPOSITION 2: *During any interval wherein investment in social infrastructure occurs and where Assumption* 1 *holds, the equilibrium condition*

$$(10) \qquad \eta W(k) = (m + r)k$$

for social capital investment per capita also holds and social capital per capita is fixed at some $k = k^*$.

PROPOSITION 3: *During any interval in which labor is used for construction and where Assumption* 1 *holds, the rate of labor use increases over time. If, in addition, Assumption* 2 *holds, during an interval when investment in social infrastructure occurs, the optimal exponential rate of labor growth* $(dL/d\text{t})/L$ is $r/(1 - h)$ *If no infrastructure investment occurs during an interval the optimal rate of labor growth is* $r/(\eta + 1 - h)$. *Thus, the rate of labor growth is greater during intervals of infrastructure investment.*

PROPOSITION 4: *If* $K^-(0) > 0$ *and assuming Assumptions* 1 *and* 2 *both hold, then labor use for construction is positive over the entire interval* $0 \leq \text{t} \leq \text{T}$.

PROPOSITION 5: *If* $K^-(0)$ *is nonnegative and sufficiently small and* $\hat{L}$, *the level of postconstruction employment, is sufficiently small, and if both Assumptions* 1 *and* 2 *hold, then there exists one point in time* t^*, $0 \leq \text{t}^* \leq \text{T}$, *such that* $I(\text{t}) > 0$ *in the interval* $0 \leq \text{t} \leq \text{t}^*$ *and* $I(\text{t}) = 0$ *in the interval* $\text{t}^* \leq \text{t} \leq \text{T}$.

PROPOSITION 6: *Given the assumptions of Propositions* 4 *and* 5, *the "final" level of cumulative investments in social infrastruc-*

ture at t^*, $K(t^*)$, *is equal to the discounted present value of the optimal net future payback of the boomtown. Thus,*

$$(11)\quad K(t^*) = K(T) = \int_{t^*}^{\infty} e^{-r(t-t^*)}[\eta WL - mK(T)]\,dt$$

where ηWL *can be considered the optimal tax collections and* $mK(T)$ *the maintenance costs at each point in time for social capital.*

II. Analysis

In this section attention will be focused on three aspects of the boom community which are of particular relevance for policy. First, we consider the nature of labor inflows to the community, and comment on some of the factors which determine the character of the boom over time. Second, we consider tax policies for financing investments in social infrastructure and the potential impacts of such policies on the time path for the construction of the new facility. Finally, we consider issues related to the optimal investment strategy for providing social infrastructure.

A. *Rates of Labor Use*

From Propositions 3, 4, and 5, percentage changes in the labor force over the construction period $0 \leq t \leq T$ are described as follows:

$$(12)\quad \frac{dL/dt}{L} = \frac{r}{1-h} \qquad \text{for } 0 \leq t \leq t^* \text{ wherein } I > 0$$

$$(13)\quad \frac{dL/dt}{L} = \frac{r}{\eta + 1 - h} \qquad \text{for } t^* \leq t \leq T \text{ wherein } I = 0$$

Thus, the character of the boom, as it relates to the rate of change in the population/labor force, is seen to depend on interest rates, the elasticity of substitution between per capita capital stocks and wages, and the structure of the production function $f(L)$. *Ceteris paribus*, the higher the interest rate the greater the boom nature of the construction period. Of course, this relationship is immediately obvious when one considers the fact that the use of labor, embodied in the partially completed (constructed) facility $C(t)$, accrues interest charges.

The dependence of $(dL/dt)/L$ on h (the elasticity of L in $f(L)$) is again obvious. If $h = 1$, in which case production is linear in L, interest charges on construction are minimized by using all required labor at one instant - the *last* instant $t = T$. For $0 < h < 1$, the larger is h the greater is $(dL/dt)/L$, and with $\bar{C}$ fixed, the more labor is used in later periods.

During the first interval $0 \leq t \leq t^*$, $I > 0$ and $dk/dt = 0$, which implies per capita social investment $k(t)$ is maintained at k^*, and $dW(k)/dt = 0$ with equality in (4). Thus $(dL/dt)/L$ is independent of η (see (12)). When investments cease and $k(t)$ begins to fall, wages must be increased to compensate for lower levels of $k(t)$ for any level of labor use by (4). The "required" percentage change in wages for any percentage change in per capita capital stock k is then given by η. Thus, the larger is η, the *slower* is the rate of increase in the labor supply. It then follows that $(dL/dt)/L$ in the investment interval $0 \leq t < t^*$ exceeds $(dL/dt)/L$ in the interval $t^* < t \leq T$.

B. *Tax Policies for Financing Social Infrastructure*

From Propositions 1 and 2, the following two conditions hold:

$$(14)\quad \eta W = (m + r)k \qquad \text{for } 0 \leq t \leq t^*$$

$$(15)\quad W(1 + \eta) = \bar{\phi} f' e^{rt} \qquad \text{for } 0 \leq t \leq T$$

As is demonstrated in Proposition 1, optimal construction rates require that the marginal value product of labor, $\phi f'$ in (15), adjusted to reflect accrued interest (e^{rt}), equals the marginal cost of labor which is the wage *plus* a charge ηW. From Assumption 1 ηW may be written as $-W'(k)k$; this can be viewed as a marginal measure of the value of reduced wages which result from providing $k(t)$ levels of

social infrastructure. Thus, ηW is essentially a measure of the rents to the construction activity (via lower wages) attributable to the existence of k. Equation (14) then requires that the marginal value product of labor cover wages plus rents. Within a decentralized setting, social optimality would then require a tax in the amount of the rents, ηW. Finally, as shown in Proposition 6, those tax collections over the life of the boomtown are sufficient to finance the optimal level of infrastructure stocks.

As is discussed in greater detail below, with rents (or taxes) assigned to the municipality, periodic investments in urban infrastructure are carried to the level where marginal tax revenues ηW are equated with marginal capital charges $(m + r)\lambda$, equation (14).

C. *Investments for Social Infrastructure*

If, as we assume to be the case, per capita capital stocks in our community at the initial moment ($t = 0$) are less than those in communities from which labor must be attracted and that such communities are in equilibrium as defined by (10) (i.e., $k(0) < k^* = \bar{k}$), there may be an initial "jump" in initial capital stocks. We then have $K^+(0) > 0$ such that per capita capital stocks in the boom community equals that in the "outside" communities, and $k(0) = k^*$. An optimal strategy then requires that investments continue over the interval $0 \leq t \leq t^*$ such that $dk/dt = 0$ (Propositions 2 and 5). Equilibrium capital stocks k^* are maintained until that period t^* at which time the value of capital stocks, $K(t^*)$, just equals the present value of net tax collections in all future periods; that is, from Proposition 6,

(16)
$$K(t^*) = \int_{t^*}^{\infty} \{\eta WL - mK(T)\} e^{-r(t-t^*)} dt$$

In (16), ηWL are optimal tax collections in all future periods (as argued above), and $mK(T)$ are periodic maintenance charges on the constant level of capital stocks, $K(T)$. The term under the integral is thus periodic *net* taxes available for financing capital stocks.

An appreciation for the nature of the decision rules governing the optimal rate of investment for social infrastructure is enhanced by the following observations. In the interval $0 \leq t \leq t^*$ during which investments take place, the following expression holds:

(17) $$W' = -(r + m) \quad \text{for } 0 \leq t \leq t^*$$

(Equation (17) is derived by differentiating Appendix equation (A7), and setting the result equal to Appendix equation (A8).) Within the context of our model, "benefits" attributable to investments in infrastructure are measured by the resulting reduction in wages required to attract the optimal level of labor. Thus, at any moment t, $0 \leq t \leq t^*$, investments are carried to the point where marginal benefits (the change—fall—in wages associated with an increment in investments) equals marginal capital costs (the interest rate plus marginal maintenance charges), as given by (17). The rule (17) is followed subject to the constraint implied by (16), viz., that marginal capital cost equals the present value of net (of maintenance costs) benefits in all future periods that result from the increment in investment. Thus, when the optimal stock $K(t^*)$ in (16) is obtained, marginal capital costs rise to the present value of future benefits and investments are terminated.

In closing this section, we wish to note that the notion of a tradeoff between W and λ has intriguing implications for a measure of considerable interest for investment planners, viz., a measure of social benefits attributable to social infrastructure. The *assumed* relation $W(k)$ may be stated as a hypothesis $W = W(k)$ which can be empirically tested. Initial efforts to do so are reported in Cummings and Mehr.

III. Concluding Remarks

In this paper we have attempted to provide an analytical framework which provides a systematic method for sorting out

the bits and pieces of the boomtown investment problem in such a way that the interrelationships between key policy variables are brought into sharp focus. The key policy issues which surface from our suggested framework include the critical interrelated nature of the timing of investments in social infrastructure and tax policies, and the role of interest rates in affecting the character of the boom.

But further, the results concerning tax policies and rates of change in L discussed above give rise to considerations directly relevant for a major set of problems facing decision makers in boomtown communities, viz., the source and timing of funds for use in investing in social infrastructure. Generally, tax revenues available to communities for such investments are derived from property and/or some kinds of ad valorem taxes. Thus, tax collections from the new facility are forthcoming only after construction activities are completed, (see Mehr, chs. 1 and 2), $t > T$, which is to say *after* the major boom period has terminated. This is the front-end problem facing boomtowns, and relates to the preinvestment activity $K^+(0)$ in which we allow a jump in capital stocks in the initial period. In our model, any level of preinvestment (front-end investments) is permitted so long as the budget constraint (16) is satisfied. Thus, one can borrow all monies which may be repaid over the interval t^* to ∞. In reality, however, most communities are faced with a number of legal as well as market restrictions on their borrowing capacity.

The ramifications of such restrictions on the community's capacity to invest are demonstrated in Figures 1 and 2. Under our optimal policy, the time path of labor use $L_1(t)$ begins at a relatively rapid rate given by (12), at t^*, the slower rate (13) applies until the completion of the construction activities at T. Per capita capital stocks ($k_1(t)$ in Figure 2) are maintained at equilibrium levels k^* until t^* at which point the budget constraint (16) becomes binding, and per capita stocks decline.

The community that lacks the funds for investments in social infrastructure begins with $k(0) < k^*$, and $I(t) = 0$ initially. It must then begin with a larger initial labor force ($L_2(0)$), and therefore a sharper initial boom impact. This follows from the fact that under optimal conditions the rate of growth in labor in the period $0 \leq t \leq t^*$ exceeds that in the community without investments (equations (14) and (15)).

As is typically the case however, if at some later period t' the community should acquire funding capacity which is used for investments in social infrastructure, two problems arise. First, boom conditions at t'

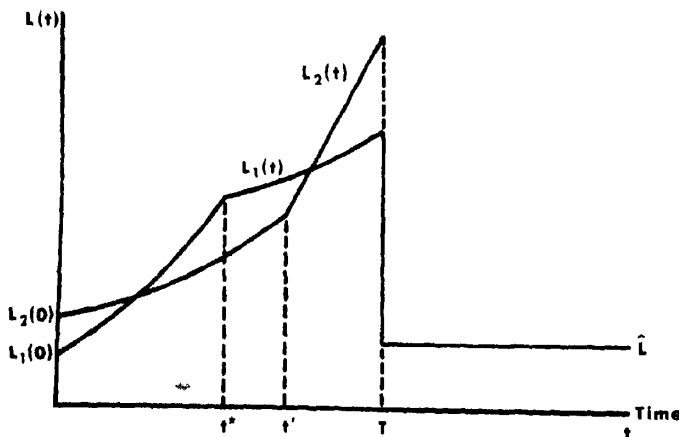



FIGURE 1

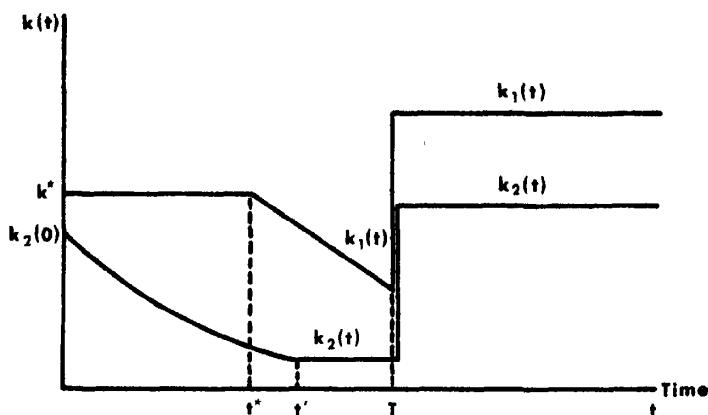



FIGURE 2

are exacerbated by an increase in the rate of change in L (moving from the rate (13) to the rate (12)) if the growth in capital (K) begins to match the growth in labor force. Second, given the shorter periods over which the tax ηW may be collected for the purpose of debt retirement, the repayable amount of investments may fall well short of equilibrium levels k^* ($k_2(t)$ in Figure 2).

These observations, of course, argue for concern as to methods for facilitating the acquisition of preboom, front-end monies by boom communities—the *timing* of such investments may be every bit as important as the amount of funds that might be eventually made available to the community. But there is a second dimension to these arguments, viz., the source of repayment funds. Optimality, within our admittedly limited framework, implies a tax related to wages as well as to individuals' elasticity of substitution of W for k. Thus, substantial tax collections are received during the construction phase. As noted above, in many communities mechanisms simply do not exist for the community to levy what one might think of as infrastructure-related "use taxes"; their major source of taxes emanate from the *completed* facility, the flow from which then begins *after* the boom period. The states receive income taxes, of course, but there is no a priori reason to expect any close relation between income taxes and the use tax ηW, the issue as to the proportion of such income taxes received by the community notwithstanding.

In terms of the tax issue, an interesting problem which should be noted but which lies well beyond the intended scope of this paper concerns the question: who pays the tax ηW? From a "use" point of view, one may argue that labor—the users of social infrastructure—should pay the tax. On the other hand, as discussed above, the community's investments in social infrastructure has the effect of generating rents to the facility, and one may argue that these rents should be taxed in order to pay for the source of such rents.

The scope of our model is certainly limited, and there are a number of potentially useful ways in which it might be extended to the end of more richly capturing the decision environment relevant for policymakers in boomtown communities. While a large number of candidates exist in terms of possible extensions of our model—for example, the multiple-boom phenomena, problems associated with imperfect information to in-migrating labor, aesthetics, and other externalities (see Ives, Schulze, and Brookshire)—we would like to comment briefly on one extension which appears to us to be of particular relevance for future empirical research concerning optimal investment strategies in boomtowns.

It would seem to be most desirable to expand the capital stock variables along two lines. First, as they relate to the welfare argument on which our analysis is based, private capital stocks are clearly relevant. By private capital stocks we refer to the availability of such things as housing and shopping centers. Second, it would be useful to differentiate between *types* of social infrastructure, for example, schools, public safety facilities (police and fire), recreation facilities, streets and roads, water supply and waste disposal facilities, etc.[2] Sharpening the focus on capital investment expenditures, a consideration from which we abstract in our model, may then require explicit treatment of the potential boom effects of such investment expenditures themselves.

Appendix: A Model of Optimal Social Investment for Boomtowns

From Section I the objective of our decision maker is to minimize the sum of social infrastructure investment costs and the labor costs of construction. Thus, we wish to minimize

$$\text{(A1)}\quad K^+(0) - K^-(0) + \int_0^T e^{-rt}[W(k)L(t) + I(t) + mK(t)]\,dt + \int_T^\infty e^{-rt}[W[K(T)/\hat{L}]\hat{L} + mK(T)]\,dt$$

subject to

$$\text{(A2)}\quad dK(t)/dt = I(t),\ I(t) \geq 0,\ 0 \leq t \leq T$$

$$\text{(A3)}\quad dC(t)/dt = f(L),\ f' \geq 0,\ f'' \leq 0,\ 0 \leq t \leq T$$

$$\text{(A4)}\quad C(T) = \bar{C}$$

Following Pontryagin et al. we form the Hamiltonian.

$$\text{(A5)}\quad H = -e^{-rt}[W(k)L(t) + I(t) + mK(t)] + \xi(t)I(t) + \phi(t)f(L)$$

where $\xi(t)$ and $\phi(t)$ are the costate variables for the state variables K and C, respectively. A necessary condition for the minimization of (A1) is that (A5) be maximized with respect to the controls L and I over the interval $0 \leq t \leq T$ where we are in effect maximizing the negative of (A1).

From Kuhn-Tucker theory, the necessary conditions for a maximum of H are as follows, given the assumption of constant utility (where $W' = dW/dk < 0$):

$$\text{(A6)}\quad \partial H/\partial L = e^{-rt}W'K/L^2 - e^{-rt}W + \phi f' \leq 0$$
$$(\partial H/\partial L)L = 0, \qquad L \geq 0$$

$$\text{(A7)}\quad \partial H/\partial I = -e^{-rt} + \xi \leq 0,$$
$$(\partial H/\partial I)I = 0, \qquad I \geq 0$$

The differential equations governing the costate variables are

$$\text{(A8)}\quad d\xi/dt = -\partial H/\partial K = e^{-rt}(W' + m)$$

$$\text{(A9)}\quad d\phi/dt = -\partial H/\partial C = 0$$

while the condition for choosing $K^+(0)$ is

$$\text{(A10)}\quad \xi(0) - 1 \leq 0$$
$$(\xi(0) - 1)(K^+(0) - K^-(0)) = 0$$
$$(K^+(0) - K^-(0)) \geq 0$$

The condition for choosing $K(T)$ is

$$\text{(A11)}\quad \xi(T) = -\int_T^\infty e^{-rt}[W' + m]\,dt$$

and the condition for the terminal constraint on construction is

$$\text{(A12)}\quad (C(T) - \bar{C})\phi(T) = 0$$
$$\phi(T) \geq 0,$$
$$(C(T) - \bar{C}) \geq 0$$

If we employ Assumption 1 we then have $(K/W)(W'/L) = -\eta$; after rearranging terms (A6)–(A12) are rewritten as follows:

$$\text{(A6}')\quad (1 + \eta)W \begin{Bmatrix} \geq \\ = \end{Bmatrix} \phi e^{rt} f' \begin{cases} \text{for } L = 0 \\ \text{for } L \geq 0 \end{cases}$$

$$\text{(A7}')\quad e^{-rt} \begin{Bmatrix} \geq \\ = \end{Bmatrix} \xi \quad \begin{cases} \text{for } I = 0 \\ \text{for } I \geq 0 \end{cases}$$

$$\text{(A8}')\quad d\xi/dt = e^{-rt}(mk - \eta W)/k$$

[2]Tentative efforts in this direction are reported in Cummings and Mehr.

(A9′) $\phi(t) = \bar{\phi} = \text{constant}$

(A10′) $\xi(0)\begin{Bmatrix} \leq \\ = \end{Bmatrix} 1 \begin{Bmatrix} \text{for } K^+(0) = K^-(0) \\ \text{for } K^+(0) \geq K^-(0) \end{Bmatrix}$

(A11′) $\xi(T) = -\int_T^{\infty} e^{-rt}[m\hat{k} - W]/\hat{k}dt$

$(k = K(T)/L)$

(A12′) $\phi(T)\begin{Bmatrix} \geq \\ = \end{Bmatrix} 0 \begin{Bmatrix} \text{for } C(T) = \bar{C} \\ \text{for } C(T) \geq \bar{C} \end{Bmatrix}$

The proofs of the propositions stated in Section I then follow immediately:

PROOF of Proposition 1:

This proposition can be shown by noting that $L > 0$ implies (A6′) holds with equality and that ϕ is a constant, $\bar{\phi}$ from (A9′). Thus we have

(A13) $(1 + \eta)W = \bar{\phi}e^{rt}f'(L)$

Since ϕ is the *present* value costate variable on construction, ϕe^{rt} is the *current* value costate variable on construction and the right-hand side of (A13) corresponds to the marginal revenue product of labor.

PROOF of Proposition 2:

With $I > 0$ equality holds in (A7′). Differentiating (A7′) with respect to time yields

(A14) $d\xi/dt = -re^{-rt}$

Setting (A14) equal to (A8′) then yields equation (10). Totally differentiating (10) with respect to t implies $dk/dt = 0$ in which case k is fixed at some $k = k^*$.

PROOF of Proposition 3:

Since $L > 0$ implies that (A13) holds, we can differentiate with respect to t obtaining

(A15) $(1 + \eta)W'\left(\frac{1}{L}dK/dt - \frac{K}{L^2}dL/dt\right) = r\bar{\phi}e^{rt}f' + \bar{\phi}e^{rt}f''dL/dt$

Rearranging terms and using Assumption 1 and (A13) we obtain

(A16) $(dL/dt)/L = \dfrac{\eta(dK/dt)/K + r}{\eta - Lf''/f'} > 0$

since $dK/dt = I \geq 0$, η, r, L, $f' > 0$, and $f'' < 0$. Thus labor use rate increases over time. Using Assumption 2 we have $f'L = hf$, which upon total differentiation yields $f''L + f' = hf'$, or $1 - h = -Lf''/f'$. Thus from (A16) we have

(A17) $(dL/dt)/L = \dfrac{\eta(dK/dt)/K + r}{\eta + 1 - h}$

If $I > 0$ then (10) holds from Proposition 2 and k is fixed at k^* so $K = k^*L$ which implies $(dK/dt)/K = (dL/dt)/L$. Thus, from (A17) and using $(dK/dt)/K = (dL/dt)/L$ we have

(A18) $(dL/dt)/L = \dfrac{r}{1 - h}$

If $L > 0$ and $I = 0$ we have

(A19) $(dL/dt)/L = \dfrac{r}{\eta + 1 - h}$

since $dK/dt = I = 0$. Finally,

(A20) $\dfrac{r}{1 - h} > \dfrac{r}{\eta + 1 - h} > 0$

since r, $1 - h$, $\eta > 0$.

PROOF of Proposition 4:

Using Assumption 2 and (A9′), condition (A6′) can be written as

(A21) $(1 + \eta)W(k) \geq \bar{\phi}e^{rt}f(L)h/L$

Since $W \propto k^{-\eta}$ Assumption 1 and $f \propto L^h$ by Assumption 2 we can show that $L \neq 0$ by noting that $\lim L \rightarrow 0\ W(K(t)/L) = 0$ since $K^-(0) > 0$ and $dK/dt = I \geq 0$ implies $K(t) > 0$, and that $\lim L \rightarrow 0\ f(L)/L = \infty$ so that the right-hand side of (A21) must be greater than the left-hand side for $L = 0$ which is a contradiction.

PROOF of Proposition 5:

(a) First consider the terminal condition on $K(T)$, (A11′), which can be rewritten by integration as

(A22) $\xi(T) = [(\eta W(\hat{k}) - m\hat{k})/\hat{k}](1/r)e^{-rt}$

since $W(\hat{k})$ and $\hat{k}$ are fixed over the interval $\infty \geq t \geq T$. If $\hat{k} > k^*$, since $W'(k) < 0$ we have by equation (10)

(A23) $[(\eta W(\hat{k}) - m\hat{K})/\hat{k}] < r$

which by (A22) implies that $\xi(T) < e^{-rT}$. This in turn implies that $I(T) = 0$ by (A7'). Note that $\hat{k} = K(T)/\hat{L}$ will be greater than $\hat{k}^*$ if $\hat{L}$ is sufficiently small and since $I(T) = 0$ we must then have $t^* < T$.

(b) Now consider the initial condition on $K(0)$, (A10'). From this we can show that if

(A24) $K^-(0) < k^*L(0)$

then

(A25) $k(0) \leq k^*$

Thus, if $I(0) > 0$ and $K^+(0) \geq K^-(0)$ then $\xi(0) = 1$ by (A10') and $k(0) = k^*$ by Proposition 2. If $I(0) = 0$, then $\xi(0) \leq 1$ by (A7') and $K^+(0) = K^-(0)$ from (A10'); thus, (A24) implies $K^-(0)/L(0) = K^+(0)/L(0) = k(0) < k^*$ where the inequality holds true if $K^-(0)$ is sufficiently small since $L(0) > 0$ by Proposition 2. It then follows that $I(0) \geq 0$ implies $k(0) \leq k^*$.

Assume that $I(0) = 0$. In this case $k(0) < k^*$ as shown in (b) above. This implies that $I(t) \equiv 0$ for $0 \leq t \leq T$ so $t^* = 0$. This can be shown by noting that for $I(t) > 0$ we must have $\xi(t) = e^{-rt}$ but that $I(0) = 0$ implies $\xi(t) < e^{-rt}$ for $0 < t \leq T$ since $\xi(0) \leq 1$ and $d\xi/dt < -re^{-rt} < 0$ for $0 \leq t \leq T$. This last inequality results from the fact that

(A26) $(mk - \eta W(k))/k < -r$

if $k(t) < k^*(t)$ where $W' < 0$ using Proposition 2. Thus, since by (A8') $d\xi/dt = e^{-rt}(mk - \eta W)/k$ we have $d\xi/dt < -re^{-rt} < 0$. $k(t)$ remains less than k^* because $k(0) \leq k^*$ if $I = 0$, and since I remains zero in the interval $0 \leq t \leq T$, and $dk(t)/dt < 0(k = k/L)$ because $(dL/dt)/L > 0$, from Proposition 3.

Now assume $I(0) > 0$. In this case $t^* < T$ as noted in (a) above. However, we can show that only one t^* exists because in any interval where $I(t) = 0$, $k < k^*$, so $d\xi/dt < -re^{-rt}$ as shown above and ξ remains less than e^{-rt}. Thus, if $I(t^*) = 0$, $I(t) = 0$ for $t > t^*$.

PROOF of Proposition 6:
Integrating (A8') yields

(A27)
$$\xi(t') - \xi(0) = -\int_0^{t'} e^{-rt}(\eta W - mk)/k\,dt$$

We then also have

(A28)
$$\xi(T) - \xi(0) = -\int_0^{T} e^{-rt}(\eta W - mk)/k\,dt$$

and subtracting (A28) from (A27) yields

(A29)
$$\xi(t') = \xi(T) + \int_{t'}^{T} e^{-rt}(\eta W - mk)/k\,dt$$

Using (A11'), (A29) becomes

(A30) $$\xi(t') = \int_{t'}^{\infty} e^{-rt}(\eta W - mk)/k\,dt$$

If we set $t' = t^*$ in (A30) and note that $K(t) = K(T)$ for $t \geq t^*$ we then have

(A31)
$$\xi(t^*) = \int_{t^*}^{\infty} e^{-rt}(\eta WL(t)/K(T) - m)dt$$

Using (A7') which implies that $\xi(t^*) = e^{-rt^*}$, multiplication of both sides of (A31) by $K(T)$ results in equation (11).

REFERENCES

R. G. Cummings, and A. Mehr, "Investments for Urban Infrastructure in Boomtowns," *Natur. Resources J.*, Apr. 1977, 223-40.

J. S. Gilmore, "Boomtowns May Hinder Energy Resource Development," *Science*, Feb. 13, 1976, *191*, 535-40.

——— **and M. K. Duff,** "The Sweetwater County Boom: A Challenge to Growth Management," work. paper series, Univ. Denver Res. Instit., July 1974.

B. Ives, W. Schulze, and D. Brookshire, "Boomtown Impacts of Energy Development in the Lake Powell Region," *Lake Powell Res. Proj. Bull.*, no. 28, Nov. 1977.

A. F. Mehr, "Measuring Social Benefits Attributable to Social Infrastructure in Boomtowns," Los Alamos Scientific Laboratory, no. LA-6559-T, Oct. 1976.

Federation of Rocky Mountain States, "Energy Development in the Rocky Mountain Region: Goals and Concerns," Denver, July 1975.

Money, Saving, and Portfolio Choice under Uncertainty

By ELIAKIM KATZ AND ALFRED VANAGS*

It is commonly held that risk aversion is an important factor in the demand for money. This view is based on the typical model of portfolio choice which takes the decision to hold a certain quantity of wealth as given and then considers the optimal distribution of wealth among different assets of varying degrees of attractiveness and riskiness (See, for example, James Tobin or M.L. Cropper, a recent paper which incorporates this approach.) In this paper we consider simultaneously the decisions on saving-borrowing and portfolio composition in the context of a simple two-period model.

We find that whether or not positive money balances are held is sensitive to the motive for saving. In particular, we show that with an additively separable intertemporal utility function, positive money balances will never be held whenever the motive for saving lies in general attractiveness of assets (albeit risky ones). We then explore more fully the structure of the joint savings-portfolio choice decision for the special case of a quadratic utility function[1] which enables the role of money to be more clearly appreciated.

We conclude that the supposition that some positive money balances will normally be held as part of an optimally diversified portfolio must be treated with care. It all depends on the motives for holding wealth. For example, if the motive for saving is to smooth out unequal endowments of wealth over time or if saving is undertaken as a precaution against uncertain future endowments rather than uncertain asset returns, positive money balances will in general reappear.

I

Consider an individual with the two-period utility function

$$(1) \qquad U = V(C_0) + kV(C_1)$$

where C_i is consumption in the ith period $i = (1, 2)$ and $k \leq 1$ is a time preference parameter. Suppose he has a certain income endowment in each period and is able to save in the first period either by accumulating riskless money m which yields a zero rate of return or by accumulating the risky asset a, the return on which is uncertain. Assuming maximization of expected utility the individual's choice problem consists of choosing a and m so as to maximize (1), that is, the individual choice problem is

$$(2) \qquad \max_{a,m} V(y_0 - (m + a)) + kEV(y_1 + m + ar)$$

where y_0 and y_1 are income endowments in the two periods, r is the (contingent) return on the a asset in the second period and E is the expectation operator. As stated, a and m are unconstrained. However we wish to interpret the risk-free asset m as money; and it seems natural to restrict holdings of m to be nonnegative. By this we wish to reflect the fact that an individual is normally unable to *sell* money short. Thus the appropriate maximand is

$$(3) \qquad L = V(y_0 - (m + a)) + kE[V(y_1 + m + ar)]$$

subject to

$$(4) \qquad m \geqq 0$$

*Lecturers, Queen Mary College, University of London. We gratefully acknowledge the helpful comments of an anonymous referee, of our colleague, Ray Rees, and also the hospitality of the University of Guelph where Vanags was a visitor during completion of the paper.

[1] The quadratic has increasing absolute risk aversion and it is well known that as a consequence there are objections to its use. However, our results do not depend on this property.

Maximizing (3) with respect to a and m we obtain the necessary conditions

$$(5)\quad \frac{\partial L}{\partial a} = -V'(y_0 - (m + a)) + kE[rV'(y_1 + m + ar)] = 0$$

$$(6)\quad \frac{\partial L}{\partial m} = -V'(y_0 - (m + a)) + kE[V'(y_1 + m + ar)] \leqq 0$$

and also the complementary slackness conditions

$$(7)\qquad m \geq 0$$

$$(8)\qquad \frac{\partial L}{\partial m} m = 0$$

We now wish to show that in a certain class of cases positive holdings of both a and m are incompatible.

Consider (5) and (6) when $y_0 \leq y_1$ and $m > 0$. From (8) $m > 0$ implies that (6) is satisfied with a strict equality, i.e.,

$$(9)\quad V'(y_0 - (m + a)) = kE[V'(y_1 + m + ar)]$$

Using the Mean Value Theorem we have that

$$(10a)\quad V'(y_0 - (m + a)) = V'(y_0) - (m + a)V''(y_0 - \theta(m + a))$$

and that

$$(10b)\quad V'(y_1 + (m + ar)) = V'(y_1) + (m + ar)V''(y_1 + \phi(m + ar))$$

where $\quad 0 \leq \theta \leq 1, \quad 0 \leq \phi \leq 1$

Substituting in (9), this yields

$$(11)\quad V'(y_0) - kV'(y_1) = (m + a)V''(y_0 - \theta(m + a)) + kE[(m + ar)V''(y_1 + \phi(m + ar)]$$

Given our assumptions, the left-hand side of (11) is clearly nonnegative. Hence, the right-hand side of (11) must be nonnegative. Since $V'' < 0$, this implies that, if $r \geq 0$, a must be negative so as to outweigh the positive m. (The condition on r will be satisfied for all limited liability risky assets such as bonds, equities, etc.)

Intuitively, the result may be seen by setting $a = 0$ in (9). Given that $y_0 \leq y_1$, and that $k \leq 1$, it is clear that when $m > 0$ this contradicts the strict equality in (9) since $V'(y_0 - m) > kV'(y_1 + m)$. Therefore, in order to have (9) as a strict equality, a must be made negative whenever m is positive, that is, the risky asset is sold, reducing marginal utility in the initial period and raising the same in the later period. Thus, under the conditions posited we have established the incompatibility of *simultaneously* positive holdings of m and a.

II

The equilibrium solution can be fully characterized for the special case where $V(C)$ is quadratic, i.e, $V(C) = C - (b/2)C^2$. In this case the first-order conditions may be written as:

$$(12)\quad 1 - b(y_0 - m - a) = k[1 - b(y_1 + m + a\bar{r})]$$

$$(13)\quad 1 - b(y_0 - m - a) = rk[1 - b(y_1 + m + a\bar{r})] - b\sigma^2 ka$$

Equations (12) and (13) yield the solution values of a and m as long as $m \geqq 0$; whenever $m < 0$ equation (12) is irrelevant and the solution for a may be obtained by setting $m = 0$ and solving for a from (13) alone.

Solving for a from (12) and (13) we obtain

$$(14)\quad a = \frac{(\bar{r} - 1)[(1 - by_0) + (1 - by_1)]}{b[(\bar{r} - 1)^2 + \sigma^2(1 + k)]}$$

From (12) we can write the solution for m as

$$(15)\quad m = \frac{(y_0 - ky_1)}{(1 + k)} - \frac{(1 - k)}{b(1 + k)} - \frac{a(1 + k\bar{r})}{(1 + k)}$$

From (15) we see that when $y_0 \leqq ky_1$, $k \leqq 1$, m will be negative whenever a is positive, that is the result we obtained above for the general utility function $V(C)$.

When $y_0 = y_1 = y$, the explicit solutions for a and m are

$$(16)\quad m = \left(\frac{1 - by}{b}\right)\frac{[1 - (1 - k)\sigma^2 - \bar{r}^2]}{[(\bar{r} - 1)^2 + (1 + k)\sigma^2]}$$

$$(17)\quad a = \left(\frac{1 - by}{b}\right)\frac{2(\bar{r} - 1)}{[(\bar{r} - 1)^2 + (1 + k)\sigma^2]}$$

whenever the solution for m is $\geqq 0$; and

$$(18)\quad a = \left(\frac{1 - by}{b}\right)\frac{(\bar{r}k - 1)}{[1 + k(\bar{r}^2 + \sigma^2)]}$$

otherwise. The implications of these conditions for the equilibrium values of m and a are summarized in Table 1.[2]

Thus, whenever the expected gross return on the risky asset is sufficiently greater than unity ($k\bar{r} > 1$) positive holdings of it will be desired, and, if it were possible, the individual would wish to sell the safe asset short. Since we interpret the safe asset as money, holdings of it are constrained to be nonnegative, yielding $m = 0$ whenever $a > 0$. Note that if the individual were unconstrained in selling m, he would choose to hold a positive quantity of a whenever $\bar{r} > 1$ (from equation (14)) and would sell the safe asset short so as to smooth out consumption over the two periods. That is, whenever $a > 0$, total saving $m + a$ would be negative for $y_0 \leq ky_1$ and $k \leq 1$, since from (15) we have

$$(19)\quad m + a = \frac{y_0 - ky_1}{1 + k} - \frac{1 - k}{b(1 + k)} - \frac{k(\bar{r} - 1)a}{1 + k}$$

which is negative under the posited conditions. When $k < 1$, an intermediate range is operative ($k\bar{r} < 1 < \sqrt{\bar{r}^2 + (1 - k)\sigma^2}$) in which a is negative and $m = 0$; or if m were unconstrained it too would be negative.

[2]We assume that y and b are such that $1 - by > 0$

TABLE 1

Condition	m	a
$1 > \sqrt{\bar{r}^2 + (1 - k)\sigma^2}$	> 0	< 0
$k\bar{r} < 1 < \sqrt{\bar{r}^2 + (1 - k)\sigma^2}$	$= 0$	< 0
$k\bar{r} > 1$	$= 0$	> 0

Finally, when the expected return on the risky asset is sufficiently low

$$(\sqrt{\bar{r}^2 + (1 - k)\sigma^2} < 1)$$

it is optimal to sell it short but *now* to hold positive quantities of m, again with the purpose of smoothing out consumption over time.

The equal endowments case thus illustrates more clearly the role of money in the savings-portfolio choice decision. When $y_0 = y_1$ (and $k = 1$) the only motive for saving lies in the attractiveness of the risky asset ($\bar{r} > 1$). If saving is undertaken, this, by definition, would drive down the expected marginal utility of consumption next period and drive up this period's marginal utility. If short sales of the safe asset were permitted they would be undertaken so as to reallocate consumption and equalize marginal utilities over the two periods. Such an operation leaves the risk borne by the individual unchanged.

REFERENCES

M. L. Cropper, "A State Preference Approach to the Precautionary Demand for Money," *Amer. Econ. Rev.*, June 1976, *66*, 388-94.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65-86.

The Golden Rule and the Role of Government in a Life Cycle Growth Model

By TOSHIHIRO IHORI*

Peter Diamond has examined the effects of government debt on long-run competitive equilibrium in a life cycle growth model. His framework is of interest because it has many features of a model of capital accumulation with decentralized decision making. Despite subsequent arguments surrounding the issuance of debt in steady-state models by A. Asimakopulos, Jerome Stein, and Paul Samuelson, the relationship between the golden rule and government debt still remains ambiguous. The purpose of this paper is to provide a useful diagrammatic exposition of some of the more obscure parts of the life cycle growth model and to clarify the implications of the argument for the role of government in the process of decentralized accumulation. The present paper argues that a long-run tradeoff between consumption possibilities plays a critical role in the discussions of long-run optimality.

In Section I, I shall offer an analysis of the long-run tradeoff between consumption in the younger generation and consumption in the older generation, and examine the significance of the golden rule in the market process. It will be shown that the golden rule path will not generally lead to a maximum of everybody's utility among all feasible steady growth paths, as long as the constraint of the market process is maintained intact. Then, in Section II, I shall analyze the effects of government debt on the long-run optimal conditions and examine the nature of the growth paths that correspond to different behavior of government. Section IIA considers Diamond's internal debt policy, and Section IIB considers an alternative policy measure, namely economic activities of government financed from debt issue. It will be shown that the desired role of government is to provide some means whereby income can be redistributed between the younger and the older generations.

I. The Golden Age Consumption-Possibility Curve

The structure of Diamond's model can be summarized as follows:

$$Y_t = F(K_t, L_t) = f(k_t)L_t \tag{1}$$

$$Y_t + K_t = K_{t+1} + e_t^1 L_t + e_t^2 L_{t-1} \tag{2}$$

$$L_t = (1+n)^t \tag{3}$$

$$L_t s_t = K_{t+1} \tag{4}$$

$$s_t = w_t - e_t^1 \tag{5}$$

$$e_t^1 + e_{t+1}^2/(1+r_{t+1}) = w_t \tag{6}$$

$$w_t = F_l = f(k_t) - f'(k_t)k_t \tag{7}$$

$$r_t = F_K = f'(k_t) \tag{8}$$

where Y_t is output, K_t is capital, L_t is labor force, e_t^1 is the younger age group's consumption per capita, e_t^2 is the older age group's consumption per capita, s_t is savings per worker, w_t is wage rate, r_t is rental price of capital, and k_t is capital per worker, subscripts designating the period.

Equation (1) represents a well-behaved aggregate production function with contant returns to scale. Equation (2) can be regarded as an equilibrium condition in the goods market. Equation (3) represents the supply of labor. Equation (4) represents the supply of capital to the capital market. Equation (5) defines each individual's savings. Equation (6) is his *ex post* budget constraint. Equation (7) can be regarded as

*University of Tokyo. I am very indebted to Kazumi Asako, Motoshige Itoh, Koichi Hamada, and Mitsugu Nakamura for their valuable comments. Thanks are also due to an anonymous referee and the managing editor. Errors and interpretations are, of course, solely my own.

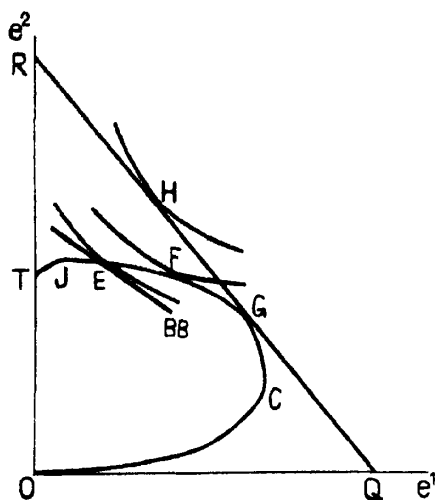



FIGURE 1

an equilibrium condition in the labor market, because exogenous labor supply (3) is fully employed through the adjustment of wage rate. Similarly, equation (8) is an equilibrium condition in the capital market In the six equations, (3)–(8), we have all the constraints of the market process.[1]

Substituting $k_t = k_{t-1} = k$ into equations (4) through (8), we have

(9) $e^1 = f(k) - [f'(k) + 1 + n]k$

(10) $e^2 = (1 + n)[1 + f'(k)]k$

These equations describe the golden age consumption possibilities (e^1, e^2) as a function of the golden age capital-labor ratio (k). In Figure 1, curve OT shows the locus of pairs of e^1 and e^2 for any level of k. It is assumed that e^1 reaches a single maximum in the range where e^2 is increasing with respect to k, and that equation (10) has at best a single maximum point. As one moves upward and to the left along curve OT, the associated capital-labor ratio becomes higher. The maximizing behavior of each individual will result in a single golden age equilibrium ratio $(\bar{k})$ on curve OT. Thus, we shall call this curve the golden age consumption-possibility curve in the market process.

If the equilibrium point E associated with $\bar{k}$ is on OC or JT where $(de^1/dk)(de^2/dk) > 0$, $\bar{k}$ is dominated by the capital-labor ratio associated with C or J in the sense that utility at C (at J) is definitely greater than utility at any point on OC (on JT). We shall call CJ where $(de^1/dk)(de^2/dk) < 0$ the efficient zone.[2] Note that from the assumption of curve OT the efficient zone is continuous.[3]

The QR line denotes the feasibility condition (2) when k is equal to the golden rule ratio k^*, so that the marginal product of capital equals the rate of growth of labor supply; hence, it represents the maximum consumption possibilities in the centrally planned economy. From equations (9) and (10), the slope of curve OT is

$$\left[\frac{de^2}{de^1}\right]_{OT} = -\frac{(1 + n)(1 + f' + kf'')}{1 + n + kf''}$$

and the slope of line QR is $-(1 + n)$. Therefore,

$$(11) \quad -\left[\frac{de^2}{de^1}\right]_{OT} - (1 + n) = \frac{(1 + n)(f' - n)}{1 + n + kf''}$$

Equation (11) implies that at G, associated with k^*, curve OT is tangent to line QR. The point G lies on the efficient zone CJ. On GJ where $k > k^*$, the slope of curve OT is algebraically less than that of line QR, and on CG where $k < k^*$, the slope of curve

[1] So long as equations (3)–(6) are satisfied, the three equilibrium conditions (2), (7), and (8) are not mutually independent If supply equals demand in both the labor market and the capital market, the equality must hold in the goods market Thus, the feasibility condition, that is, the equilibrium condition (2) will not be considered explicitly

[2] We focus on a comparison of golden ages, not on a change from one to the other Discussions of dynamic efficiency including the problem of initial condition are ignored in this paper

[3] If we assume that equations (9) and (10) are strictly concave with respect to k, curve OT is strictly convex to the right. For example, if $f(k) = k^\beta$ $(0 < \beta < 1)$, this condition is satisfied. However, curve OT need not be convex. All we require is that the efficient zone CJ is continuous.

OT is algebraically greater than that of line QR. Except at G, the consumption possibilities in the market process are more restricted than in the centrally planned economy. In a competitive economy the capital market represents an additional constraint besides the feasibility condition, namely a decentralized arrangement in which savings by the younger generation are used to provide capital for the older period.

The utility maximizing behavior of each individual of the younger generation implies that equations (5) and (6) are replaced by his optimal saving behavior; that is,[4]

$$\text{(12)} \qquad \text{Max } U(e^1_t, e^2_{t+1})$$

subject to $e^1_t + e^2_{t+1}/(1 + r_t) = w_t$

The quantity saved will be expressed as a function of w_t and r_t:

$$\text{(13)} \qquad s_t = s(w_t, r_t)$$

From the five equations (3), (4), (7), (8), and (13), we obtain the competitive growth path in the market process. It is assumed that the economy has a single stable long-run equilibrium point

Since the budget constraint can be drawn at each point on curve OT, maximization of (12) implies that the competitive point E is given where the slope of the budget line BB equals the slope of an indifference curve. In general, E does not coincide with G. When each individual has a utility function such that the competitive maximization problem (12) does not lead to the golden rule path, the utility level at E is less than at H, where line QR is tangent to an indifference curve. When E is not G, E cannot coincide with H. Thus, we can summarize the significance of the golden rule as follows: (a) If a steady-state path is not the golden rule path, it is inefficient in the sense that everybody's utility could be increased by interfering with individual maximizing behavior.

It is well known that in general the golden rule has another normative meaning: (b) The golden rule path maximizes everybody's utility among all feasible steady paths. In the present case, however, because H does not exist on curve OT, each individual's utility cannot be increased to the utility level at H even if the golden rule path G is realized. To realize the utility level at H, curve OT needs to be altered by the intervention of government. As long as all the restraints in the market process (i.e., equations (3)-(8)) are maintained intact, the optimal point is F associated with $\hat{k}$, where curve OT is tangent to an indifference curve. Since the utility level at F is greater than at G but is less than at H, we shall call F the second best point.[5] We now find that the normative meaning (b) does not always hold in the market process.

Obviously, F exists on the efficient zone CJ. We have the following proposition:

PROPOSITION:[6] *If the competitive capital-labor ratio is not equal to the golden rule capital-labor ratio, the second best capital-labor ratio exists between them. In other words,*

$$\text{If } \bar{k} \gtreqless k^*, \text{ then } \bar{k} \gtreqless \hat{k} \gtreqless k^*$$

PROOF:

(i) It is obvious that if $\bar{k} = k^*$, then $\bar{k} = \hat{k} = k^*$. (ii) If $\bar{k} > k^*$, then the marginal rate of substitution (MRS) at E equals $1 + f'(\bar{k}) < (1 + n)$. Since MRS at G < MRS at E, MRS at $G < 1 + n$. It means that at G the slope of an indifference curve is algebraically greater than the slope of curve OT. Hence, $\bar{k} > \hat{k} > k^*$. (iii) If $\bar{k} < k^*$, now suppose $k^* < \hat{k}$ (see Figure 2).

[4] Maximization of (12) implies that each individual expects r_{t+1} to be equal to r_t. On a golden age path, each individual will consume exactly the desired amount in the older period that he plans in the younger period.

[5] To realize the second best growth path, the following procedures can be considered. In one procedure, each individual maximizes his utility subject to all the constraints; he can determine the amount of his savings at the level of $(1 + n)\hat{k}$. As an alternative, the government influences the maximizing behavior (12) through some noneconomic procedure such as propaganda to increase savings. The government may control savings so that the long-run capital-labor ratio equals $\hat{k}$.

[6] This problem was discussed by Stein.

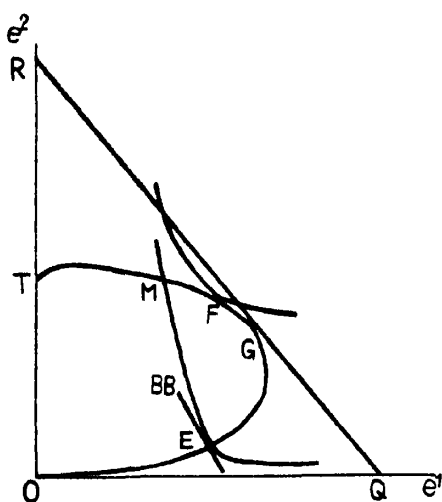



FIGURE 2

Then at F, $MRS < 1 + n < 1 + f'(\bar{k})$, and at M where an indifference curve I_0 crosses curve OT, $MRS > 1 + f'(\bar{k})$. Since MRS is a continuous function of the amount of goods consumed, there exists point Z between F and M such that MRS at Z equals $1 + f'(\bar{k})$. This means that the budget constraint just touches the indifference curve at Z as well as at E. From the assumption of the uniqueness of the competitive solution, Z is E, a contradiction.

REMARK: As shown in Figure 1, when F is on GJ (on CG), H is on GR (on QG). Thus, we know that if $\bar{k} > k^*$, H exists on GR (if $\bar{k} < k^*$, H exists on QG). To realize the utility level at H, when $\bar{k} > k^*$ curve OT needs to be shifted upward and to the left, and vice versa, by the intervention of government.

II. The Role of Government

We shall analyze the intervention of government to shift curve OT, and examine the resulting optimal growth path. We shall assume that the government issues debt to the younger generation. This debt has a one-period maturity and will be repaid with interest at the same rate of return on capital in the next period. The government ensures that a constant amount of debt per worker (g) is maintained.

A. *Transfer Program*

Here it is assumed that the government does not invest the amount borrowed or raised through taxation, but uses the funds to pay existing debt with interest. Let us denote the constant lump sum tax levied on the younger generation and the older generation by t_1 and t_2, respectively. Then we have

$$(14) \quad gL_{t-1}(1 + r_t) - gL_t = t_1 L_t + t_2 L_{t-1}$$

Each individual's disposable income ($\hat{w}_t$) is given by $(w_t - t_1 - g)$, his disposable income in the younger period t plus $(g - t_2/(1 + r_t))$ the present value of debt redemption and tax in the older period t + 1. Note that g can be negative, in which case g means "negative debt"; that is, the government lends g to each individual of the younger generation and will redeem this debt with interest.

$$(15) \quad \hat{w}_t = w_t - t_1 - t_2/(1 + r_t)$$

Let us define the relative burden ratio v by

$$(16) \quad v = t_2/t_1(1 + n)$$

Then, from equations (14), (15), and (16), we have

$$(17) \quad \hat{w}_t = w_t - \frac{[(1 + n)v + (1 + r_t)](r_t - n)}{(1 + r_t)(1 + v)(1 + n)} \cdot g$$

Now, the maximizing behavior of each individual is represented by maximization of (12), substituting $\hat{w}_t$ for w_t. However equation (4) must be altered, because each individual in fact lends $g - t_2/(1 + r_t)$ to the government on the same terms as funds could be lent in the capital market. Hence we have

$$(18) \quad L_t\left(s_t - g + \frac{t_2}{1 + r_t}\right) = K_{t+1}$$

By substituting $\hat{w}_t$ for w_t in (5) and (6), retaining (3), (7), and (8), and replacing (4) by (18), we have all the restraints in the presence of the debt issue. Then in a golden age we have

$$(19) \quad e^1 = f(k) - (1 + n + f')k - \frac{1 + f' + v(1 + n)}{(1 + n)(1 + v)}g$$

$$(20) \quad e^2 = (1 + n)(1 + f')k + \frac{1 + f' + v(1 + n)}{1 + v}g$$

The two equations (19) and (20) imply that the golden age consumption-possibility curve now depends on g and v as well as k. The last terms

$$\left[-\frac{1 + f' + v(1 + n)}{(1 + n)(1 + v)}g\right]$$

and

$$\left[\frac{1 + f' + v(1 + n)}{1 + v}g\right]$$

reflect transfer programs of the government; the government collects the amount of

$$\frac{1 + f' + v(1 + n)}{(1 + n)(1 + v)}g$$

per capita from the younger generation and transfers it to the older generation. When $g > 0$ ($g < 0$), curve OT will shift upward and the left (downward and to the right). Therefore, in Figure 3 on the left where H is on GR, g needs to be positive, and in Figure 3 on the right where H is on QR, g needs to be negative.

Since maximization of (12) holds, the slope of the indifference curve at the long-run competitive capital-labor ratio $\bar{k}(g, v)$ equals the slope of the budget constraint $-(1 + f'(\bar{k}))$, while the slope of the indifference curve at H equals $-(1 + n)$. Thus, the optimal level of g is such that $\bar{k}(g, v)$ is equal to k^*.

Considering equation (18), $\bar{k}(g, v)$ is determined by

$$(21) \quad s(\hat{w}, r) = (1 + n)k + \frac{1 + f' + v(1 + n)}{(1 + f')(1 + v)}g$$

Taking the total derivative of k with respect to g, we have

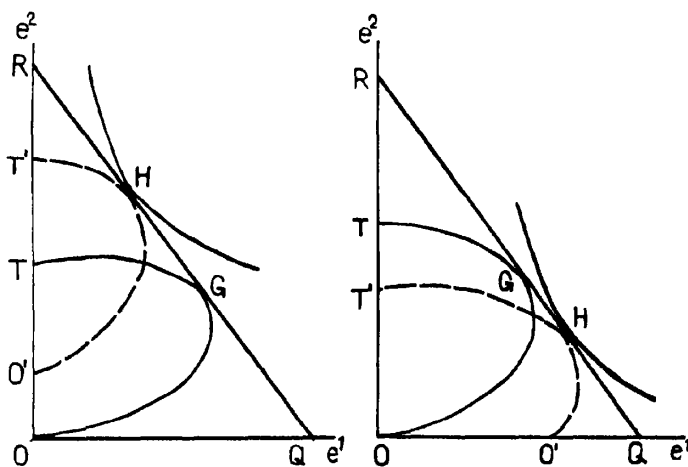



FIGURE 3

$$(22)\quad \frac{dk}{dg} = \frac{\dfrac{\partial s}{\partial g} - \dfrac{1 + f' + v(1 + n)}{(1 + f')(1 + v)}}{1 + n - \dfrac{\partial s}{\partial k}}$$

From the assumption of the stability of the system the denominator is positive, and from the assumption of the normality of the utility function the numerator is negative.[7] Therefore, dk/dg is definitely negative. The utility of each individual living in the long run will be increased by increasing g whenever $\bar{k}(g, v) > k^*$, and by decreasing whenever $\bar{k}(g, v) < k^*$. The government can shift curve OT so that H is just associated with the competitive capital-labor ratio $\bar{k}(g, v)$.[8]

The following cases are of considerable interest:

$v = 0$, i.e., $t_2 = 0$: The tax collected to finance interest costs will be lump sum taxes on the younger generation. This debt issue corresponds to Diamond's internal debt.

$v = \infty$, i.e., $t_1 = 0$: This debt issue is equivalent to tax-financed transfer payments. The government levies the lump sum tax g on the younger generation and transfers it to the older generation in the same period.

The tax-financed transfer payments and Diamond's internal debt have the same effect on the long-run equilibrium. In other words, this national debt can be regarded as a device which is used to redistribute income between the younger and the older generations.

B. *Economic Activities of Government*

We shall now examine economic activities of government. Government capital investment will be financed from debt issue. We shall assume that the technology of government production is the same as that of private production. Since receipts from the debt issue in period t will be invested in production as government capital in period t + 1, we have

$$(23)\qquad L_t g = L_{t+1} k_{g,t+1}$$

where $k_{g,t}$ denotes per worker government capital in period t. Rewriting per worker private capital in period t as $k_{p,t}$, per worker total capital in period t is defined by

$$(24)\qquad k_t = k_{g,t} + k_{p,t}$$

Because of economic activities, the government need not levy taxes; that is, $t_1 = t_2 = 0$. Hence, the disposable income is $w_t - g + g = w_t$, and the budget constraint (6) will be unchanged. By substituting $k_{p,t}$ for k_t in (18), retaining (3) and (5)-(8) and adding (23) and (24), in a golden age we have equations (9), (10), and

$$(25)\qquad k = k_p + g/(1 + n)$$

Thus, the consumption-possibility curve corresponding to government capital is nothing more than curve OT. It is true that the long-run private capital-labor ratio $\bar{k}_p(g)$ depends on g, but the long-run total capital-

[7]As shown by Diamond, the assumption of the stability of the system means $(1 + n) - \partial s/\partial k > 0$, and the assumption of the normality of the utility function means $0 < \partial s/\partial \hat{w} < 1$. The numerator of equation (22) may be written as

$$A = \frac{\partial s}{\partial g} - \frac{1 + f' + v(1 + n)}{(1 + f')(1 + v)}$$

$$= \frac{\partial s}{\partial w}\,\frac{[(1 + n)v + 1 + f'](n - f')}{(1 + f')(1 + v)(1 + n)} - \frac{1 + f' + v(1 + n)}{(1 + f')(1 + v)}$$

$$= \frac{(1 + n)v + 1 + f'}{(1 + f')(1 + v)}\left[\frac{n - f'}{1 + n}\cdot\frac{\partial s}{\partial \hat{w}} - 1\right]$$

Considering the normality of the utility function, we we know when $f' > n$, $A < 0$, when $f' < n$,

$$A < \frac{(1 + n)v + 1 + f'}{(1 + v)(1 + f')}\left[\frac{n - f'}{1 + n} - 1\right] = -\frac{[(1 + n)v + 1 + f'](1 + f')}{(1 + v)(1 + f')(1 + n)} < 0$$

[8]The relative burden ratio v is assumed to be exogenously given as a constant. Similarly, changes in v can be examined as shown by G. O. Bierwag et al. However, since v cannot be negative, efficiency will be restricted in this case.

labor ratio is independent of g.[9]

This means that public investment would exactly replace private investment. In other words, contrary to the usual supposition, the long-run effects of public investment are null. What is the reason for this paradoxical result? It is simply that the government neither levies a tax nor grants a subsidy, and promises to pay the same interest rate on the debt as could be earned on the purchase of real capital. Since the budget constraint is expressed by the same equation in the market process, the government cannot redistribute income between generations. The amount of each individual's savings determined by maximization of (12) does not depend on g.

However, once elements of private myopia are introduced into the model, it can be shown that government economic activities can increase each individual's utility in the long run. It is now assumed that by some myopic elements each individual takes his disposable income as $w_t - g$, without anticipation of the receipts from government production in the later period.[10] Hence, substituting $w_t - g$ for w_t in equation (13) and considering a modified equation (18), the golden age competitive capital-labor ratio $(\bar{k})$ is given by

$$(26) \qquad s(w - g, r) = (1 + n)k - g$$

Taking the total derivative of k with respect to g, we have

$$(27) \qquad \frac{dk}{dg} = \frac{1 + \dfrac{\partial s}{\partial g}}{1 + n - \dfrac{\partial s}{\partial k}}$$

As in equation (22), from the assumption of stability of the competitive solution the denominator is positive, and from the assumption of normality of the utility function the numerator is also positive. Any increase in public spending has an expansionary long-run effect in this case. However, substituting $w_t - g$ for w_t in equation (5) and retaining equations (3), (6)–(8), and (18), we know that the golden age consumption-possibility curve is still curve OT. The government can influence the golden age competitive equilibrium, but cannot shift curve OT. Therefore, the target point is the second best point F. Equation (27) means that the government can realize the second best path by increasing g, provided F exists on CG in Figure 1.[11]

III. Concluding Remarks

On a golden age path, the amount that each member of any given generation will consume in each period of his lifetime is uniquely determined by the long-run capital-labor ratio. Curve OT implies that consumption possibilities are more restricted than in the centrally planned economy.

In the case of Diamond's internal debt, the optimal path is always the golden rule path, because the government can shift curve OT in order to redistribute income between generations as in the centrally planned economy.

It should be stressed that the productive effect of government activities financed by debt has little effect on the long-run total capital-labor ratio. The government cannot shift curve OT so that the target path will remain the second best path, the same as in the market process. This suggests that the desired role of government in the present framework is to provide some means whereby income can be redistributed between the younger and the older generations.[12]

[9] From equation (25), obviously

$$\frac{dk_p}{dg} = -\frac{1}{1+n}, \text{ and } \frac{dk}{dg} = 0$$

[10] It is interesting to examine the myopic effect ignored by the above discussions, although there may be nothing that is not anticipated in the long run.

[11] One could say that the meaning of public investment is, at most, noneconomic propaganda. In other words, the myopic elements might reflect the noneconomic propaganda by government. See fn. 5.

[12] However, if we consider the international capital market, public investment retains an important role. In such a case the government could help to realize a higher utility level than on the golden rule path through capital export or import.

REFERENCES

A. Asimakopulos, "The Biological Interest Rate and the Social Utility Function," *Amer. Econ. Rev.*, Mar. 1967, *57*, 185–89.

G. O. Bierwag, M. A. Grove, and C. Khang, "National Debt in a Neoclassical Growth Model: Comment," *Amer. Econ. Rev.*, Mar. 1969, *59*, 205–10.

P. Diamond, "National Debt in a Neoclassical Growth Model," *Amer. Econ. Rev.*, Dec. 1965, *55*, 1126–50.

P. Samuelson, "Optimum Social Security in a Life-Cycle Growth Model," *Int. Econ. Rev.*, Oct. 1975, *16*, 539–44.

J. L. Stein, "A Minimal Role of Government in Achieving Optimal Growth," *Economica*, May 1969, *34*, 139–50.

Vertical Integration, Tying, and Antitrust Policy

By ROGER D. BLAIR AND DAVID L. KASERMAN*

A recent series of articles by John Vernon and Daniel Graham, Richard Schmalensee, George Hay, and Frederick R. Warren-Boulton has demonstrated that an input monopolist selling to a competitively structured downstream industry cannot reap the full monopoly rents available in the final-product market when substitution possibilities are present in the final-goods production function (i.e., under variable proportions). As a result, intermediate-product monopolists have been shown to have an incentive to integrate downstream, possibly engaging in a price-cost squeeze of the final-good producers in an attempt to extend the monopoly to the terminal stage of production and obtain the maximum profits available in the industry under the cost and demand conditions extant.[1] Although the variable proportions incentive for vertical integration had been explored in an earlier article by Meyer Burstein,[2] its validity and relevance for antitrust enforcement were not generally recognized until formal proof was provided.

The original Burstein article, however, also argued that the upstream monopolist could obtain identical results by tying the purchase of nonmonopolized substitutable inputs to the purchase of the intermediate product over which the monopolist exercises control.[3] If this proposition is true, then an intermediate-product monopolist has an alternative to the strategy of vertical integration which may be employed in circumstances where ownership integration is either infeasible or unattractive. More importantly, this proposition implies a previously unrecognized symmetry between the economic effects of tying and vertical integration, which in turn raises serious questions regarding the divergent treatment afforded these alternative strategies under existing antitrust laws. The purpose of this note is to provide a formal proof of the Burstein proposition and to briefly explore the antitrust policy issues that are thereby raised.

I. Equivalence of Tying and Vertical Integration

We adopt the following notation and assumptions: $P(Q)$ = final-product inverse

*Associate professor of economics, University of Florida, and economist, the Federal Trade Commission, respectively. Although this paper has benefited from the cogent comments and criticisms of several colleagues, Yoram Peles, David Qualls, John Siegfried, and an anonymous referee, we assume responsibility for what follows. Financial support of the Public Policy Research Center at the University of Florida was most welcome in the early stages. The final work was done while Blair was at the Center for the Study of American Business at Washington University.

[1]The incentive to integrate vertically has recently been generalized to the case of imperfect competition at one stage of a vertical market chain by Martin Perry. Since imperfectly competitive behavior on one side of an intermediate product market leads to an undervaluation of the properties of competitive producers on the other side of the market, the assumption of pure monopoly at the upstream stage does not appear to be crucial to the analysis.

[2]See Burstein's paper. Also, the efficiency implications of variable proportions under input monopoly were noted briefly (but not explored in the context of vertical integration) by Lionel McKenzie.

[3]Such a tying arrangement is labeled as a "full-line force" by Burstein. This terminology, however, implies that downstream producers are required to purchase a complete line of production-related (perhaps joint) intermediate products from the upstream monopolist's output vector. Instead, as will be seen below, the producer of the tying (monopolized) good need not manufacture the tied good(s) at all. Rather, he may purchase them from other producers and resell them to his downstream customers. Consequently, the discussion here will make use of the more general terminology of tying arrangements. See Hay for a restatement of the basic full-line forcing proposition.

demand; $Q(X_1, X_2)$ = final-product production function, assumed linearly homogeneous in inputs X_1 and X_2; and C_i = constant marginal cost of input X_i, $i = 1, 2$.

Suppose that the production of X_1 is monopolized while the markets for X_2 and Q are competitive. Then, if $Q(X_1, X_2)$ admits variable proportions, full monopoly rents cannot be obtained solely from the sale of X_1, because the derived demand for the monopolized input will reflect both consumer and downstream producer substitution in response to a supracompetitive input price. Producer substitution leads to economically inefficient production, which the input monopolist can circumvent through forward integration.[4]

Suppose that such integration results in successful monopolization of the market for Q.[5] Then, adoption of this strategy yields the following profit function for the integrated monopolist:

$$(1) \quad \pi_I = P[Q(X_1, X_2)] \cdot Q(X_1, X_2) - C_1 \cdot X_1 - C_2 \cdot X_2$$

since X_1 is priced internally at marginal cost.

Alternatively, the intermediate-good monopolist can purchase X_2 at the competitive price C_2 and tie its purchase by competitive downstream producers to the purchase of the monopolized input X_1. Following this strategy, the firm has a profit function:

$$(2) \quad \pi_T = p_1(X_1, X_2) \cdot X_1 + p_2(X_1, X_2) \cdot X_2 - C_1 \cdot X_1 - C_2 \cdot X_2$$

where p_i is the price of the ith input to the competitive downstream producers.

Under these conditions, vertical integration and tying are economically equivalent. That is, these strategies yield identical results with regard to both profitability and productive efficiency. This assertion can be shown by proving the following two propositions.

PROPOSITION 1: *Given a monopoly in X_1, vertical integration that results in a monopolization of the market for Q and tying the purchase of X_2 to the purchase of X_1 yields identical profits to the input monopolist.*

PROOF:

We want to show that $\pi_I = \pi_T$. Cancelling input costs from (1) and (2), this requires that

$$(3) \quad P[Q(X_1, X_2)] \cdot Q(X_1, X_2) = p_1(X_1, X_2) \cdot X_1 + p_2(X_1, X_2) \cdot X_2$$

Under the tying arrangement, the competitive downstream producers accept p_1, p_2, and P as given. Profit maximization by these firms requires that the value of the marginal product of each input be equated to its price, i.e.,

$$(4) \quad P[Q(X_1, X_2)] \cdot \frac{\partial Q}{\partial X_i} = p_i(X_1, X_2) \qquad i = 1, 2$$

Substituting (4) into (3) and factoring P, we obtain

$$(5) \quad P[Q(X_1, X_2)] \cdot Q(X_1, X_2) = P[Q(X_1, X_2)] \cdot \left(\frac{\partial Q}{\partial X_1} \cdot X_1 + \frac{\partial Q}{\partial X_2} \cdot X_2\right)$$

which yields the desired result

$$(6) \quad P[Q(X_1, X_2)] \cdot Q(X_1, X_2) = P[Q(X_1, X_2)] \cdot Q(X_1, X_2)$$

by Euler's theorem.[6] This establishes the

[4] This is, of course, the result that is due to Vernon and Graham, Schmalensee, Hay, and Warren-Boulton, which was anticipated by Burstein a decade earlier.

[5] See Warren-Boulton for a description of the circumstances that would facilitate this sort of monopoly extension. In essence, the downstream production function must exhibit substitution possibilities, while at the same time disallow production with a zero input of the monopolized intermediate good. That is, $Q(X_1, X_2)$ must be strictly quasi concave, and $Q(0, X_2) = 0$.

[6] Even if the production function is not everywhere linearly homogeneous, at a point of competitive equilibrium, where average cost is at a minimum it is locally true that $x_1 \cdot \partial Q/\partial x_1 + x_2 \cdot \partial Q/\partial x_2 = Q$.

profitability equivalence of the alternative strategies. It remains to state and prove the second proposition:

PROPOSITION 2: *Inputs X_1 and X_2 will be employed in efficient proportions whether the monopolist obtains control of the final-good industry through vertical integration or engages in a tying arrangement.*

PROOF:
Efficient production requires that input proportions be adjusted such that

$$\frac{\partial Q}{\partial X_1} \Big/ \frac{\partial Q}{\partial X_2} = \frac{c_1}{c_2} \tag{7}$$

that is, that the firm be on its expansion path. The first-order conditions for profit maximization by the integrated monopolist are

$$\left(P + Q\cdot\frac{\partial P}{\partial Q}\right)\cdot\frac{\partial Q}{\partial X_1} = c_1 \tag{8}$$

$$\left(P + Q\cdot\frac{\partial P}{\partial Q}\right)\cdot\frac{\partial Q}{\partial X_2} = c_2 \tag{9}$$

Dividing (8) by (9) yields expression (7), which establishes efficient input utilization under vertical integration. To establish the equivalent result under a tying arrangement, substitute (4) into (2) and differentiate with respect to X_1 and X_2. This yields the first-order conditions for a maximization of π_7:

$$\left[P + \frac{\partial P}{\partial Q}\cdot\frac{\partial Q}{\partial X_1}\cdot X_1 + \frac{\partial P}{\partial Q}\cdot\frac{\partial Q}{\partial X_2}\cdot X_2\right]\cdot\frac{\partial Q}{\partial X_1} + P\cdot\left[X_1\cdot\frac{\partial^2 Q}{\partial X_1^2} + X_2\cdot\frac{\partial^2 Q}{\partial X_2 \partial X_1}\right] = c_1 \tag{10}$$

$$\left[P + \frac{\partial P}{\partial Q}\cdot\frac{\partial Q}{\partial X_1}\cdot X_1 + \frac{\partial P}{\partial Q}\cdot\frac{\partial Q}{\partial X_2}\cdot X_2\right]\cdot\frac{\partial Q}{\partial X_2} + P\cdot\left[X_2\cdot\frac{\partial^2 Q}{\partial X_2^2} + X_1\cdot\frac{\partial^2 Q}{\partial X_1 \partial X_2}\right] = c_2 \tag{11}$$

Linear homogeneity of the production function implies that the second bracketed term on the left-hand side of (10) and (11) is zero.[7] Dropping these terms, we factor $\partial P/\partial Q$ from the last two terms of the first bracketed expression, apply Euler's theorem, and write conditions (10) and (11) as

$$\left(P + Q\cdot\frac{\partial P}{\partial Q}\right)\cdot\frac{\partial Q}{\partial X_1} = c_1 \tag{12}$$

$$\left(P + Q\cdot\frac{\partial P}{\partial Q}\right)\cdot\frac{\partial Q}{\partial X_2} = c_2 \tag{13}$$

Division again results in expression (7), thereby establishing efficient input utilization under a tying arrangement.[8]

Propositions 1 and 2 establish a symmetry between both the private and the social effects of vertical integration and tying under input monopoly and variable proportions. Given such symmetry, the firm holding monopoly power over an input for which (imperfect) substitutes exist must select between the alternative strategies on the basis of factors that lie outside the simplified model employed above. Important considerations that should influence this choice include: the number of substitutable inputs that must be tied to the monopolized input in order to ensure efficient down-

[7]This follows from a property of linearly homogeneous functions, viz.,

$$\frac{\partial^2 Q}{\partial x_1^2} = -\frac{X_2}{X_1}\frac{\partial^2 Q}{\partial X_2 \partial X_1}$$

See R. G. D. Allen, pp. 315-22, for a discussion of homogeneous functions.

[8]In effect, optimal use of the tying alternative will result in the downstream producers' return to the expansion path that competitively determined prices would generate. In our model, this requires that the monopolist adjust the prices of all tied inputs such that the *relative* prices remain the same as competitively determined relative prices (i.e., all input prices must be adjusted above marginal cost in equal proportion). Alternatively, the monopolist could determine the precise input proportions necessary for production on the expansion path and tie the inputs in those proportions. This may not be too difficult for homogeneous production functions since the expansion path would be linear and optimal input proportions would be independent of scale of output. But for the general case, the task could be formidable.

stream production;[9] potential cost savings that may be available through vertical integration because of transactional efficiencies or technological inseparability of the various stages of production;[10] and the comparative treatment afforded the alternative strategies by the antitrust authorities. The first two factors relate to the relevant characteristics of the markets for intermediate products and the existing production technology. The third, however, is determined by the policy thrust of the enforcement agencies. This aspect of the institutional framework will be examined in the next section.

II. Asymmetric Treatment of Symmetric Practices

To the extent that vertical integration and tying receive divergent treatment by the antitrust authorities and the courts, there will exist a bias in the firm's selection process in favor of the safest strategy. Under existing policies, this bias would appear to encourage firms to employ the strategy of vertical integration, particularly in situations where such integration can be carried out by internal expansion. In these situations, there is little chance that the practice will be challenged at all until concentration at the downstream stage approaches the monopoly level. And then, the case would be subject to a rule of reason treatment under Section 2 of the Sherman Act

Vertical integration by acquisition or merger would appear to be considerably more hazardous. While a rule of reason approach would still be applied, Section 7 of the Clayton Act would be brought to bear.[11] In such cases, the evidential burden placed on the enforcement agency is reduced and the practice is open to attack in its incipiency. Therefore, the risk of prosecution may be much higher if forward integration is carried out in this fashion. But, more important for our purposes here, it is likely that such risk will still be lower than that faced by the firm opting for the tying arrangement alternative.

Precedent indicates that a tying arrangement will be found per se illegal under Section 1 of the Sherman Act if: (a) the tying product provides the supplier with sufficient economic power to control prices; and (b) a not insubstantial dollar volume of commerce in the tied product has been affected. And if either condition holds, Section 3 of the Clayton Act may be violated.[12] The requirement of finding "sufficient economic power" has been weakened over time. In *Northern Pacific Railway Co. v. United States*,[13] such power was found to exist in the fact that the railroad's land, the tying good, was uniquely and strategically located In *United States v. Loew's Inc.*, sufficient economic power was inferred from the fact that the tying products were copyrighted.[14] Finally, in *Fortner Enterprises, Inc v United States*, the Supreme Court

[9]The franchising phenomenon may be seen as a method of assuring that inputs will be employed in efficient proportions when the number of such inputs is relatively large. Thus, some of the contractual provisions imposed upon franchisees amount to nothing more nor less than input tying.

[10]These potential cost savings are discussed in some detail by Oliver Williamson (1971)

[11]Generally, a vertical merger can be challenged on market foreclosure grounds. For our purposes, however, this is not relevant because the input monopolist has no obvious competitors that could be foreclosed from access to the final-goods producers. However, arguments that rely on the potential competition doctrine or accusations that the input monopolist is insulating its monopoly position by raising entry barriers through acquisition of its customers could be brought to bear

[12]This distinction with respect to standards of proof was made clear in *Times-Picayune Publishing Co v United States*, "When the seller enjoys a monopolistic position in the market for the 'tying' product, *or* if a substantial volume of commerce in the 'tied' product is restrained, a tying arrangement violates the narrower standards expressed in Section 3 of the Clayton Act because from either factor the requisite potential lessening of competition is inferred. . . .a tying arrangement is banned by Section 1 of the Sherman Act whenever *both* conditions are met" (pp. 608-09).

[13]The railroad tied its transportation services to the sale or lease of its land.

[14]This case involved the practice of block booking the television exhibition rights to copyrighted films. For an interesting, and succinct analysis of the block booking phenomenon, see George J Stigler

originally found that one could infer the requisite economic power from particularly advantageous terms and prices.[15] Also, establishing that a not insubstantial dollar volume of commerce has been restrained does not appear difficult. In *Fortner. . .*, $180,000 was found to be not insubstantial as was $60,800 in *Loew's. . . .* Although there are exceptions to the rather simplistic view that tying arrangements are illegal per se, the bulk of the enforcement experience indicates that it is far more hazardous than vertical integration.

III. Public Policy Implications

We have demonstrated that vertical integration and tying arrangements are alternative means of obtaining precisely the same results for an input monopolist facing variable proportions at the downstream stage For purposes of ultimate public policy in an absolute sense, we should note that the welfare implications of either strategy are ambiguous Both result in an improvement in the efficiency of downstream production because the socially optimal combination of inputs will be used instead of a nonoptimal combination that would be selected if the monopolist were required to extract rents from the single intermediate product. At the same time, there is a restriction of output to the monopoly level. Thus, the net welfare effects are indeterminate on an a priori basis, as noted by Warren-Boulton.[16] But on the one hand, forward integration by internal growth or merger receives a rule of reason treatment under either Section 2 of the Sherman Act or Section 7 of the Clayton Act. On the other hand, tying arrangements are treated quite severely as per se violations of Section 1 of the Sherman Act or Section 3 of the Clayton Act. This distinction is important as it affects the managerial decision process in selecting between these alternative strategies. The danger, of course, is that an equivalent economic result may be obtained at a higher social cost if antitrust enforcement biases the managerial decision in favor of vertical integration.[17] Thus, in the context that we have analyzed, vertical integration and tying deserve symmetric treatment under the antitrust laws; both should be judged according to the rule of reason.

[15]It was alleged that U S Steel conditioned the provision of loans on extremely desirable terms to Fortner's purchase of prefabricated homes produced by U S Steel Thus, the more competitive a firm is the more likely the Court was to find the presence of economic power. In a subsequent (and presumably final) decision in *Fortner*, the Court cleared up some of this illogic "Quite clearly, if the evidence merely shows that credit terms are unique because the seller is willing to accept a lesser profit or to incur greater risks than its competitors, that kind of uniqueness will not give rise to any inference of economic power in the credit market" *United States Steel Corporation v. Fortner Enterprises, Inc.*, p 4174.

[16]In addition, see Williamson (1968) for an analysis of the antitrust implications of a juxtaposition of monopoly power and productive efficiency.

[17]There is always some concern that vertical integration by merger or by internal expansion passes a kind of market test while tying arrangements are agreements that bypass the market In a situation where one of the inputs is monopolized, however, the concern is not well founded Forcing the input monopolist to vertically integrate in order to exploit fully his monopoly merely serves to require that the profit foregone due to input substitution exceed the added costs of vertical integraion

REFERENCES

R. G. D. Allen, *Mathematical Analysis for Economists*, New York 1938.

M. L. Burstein, "A Theory of Full-Line Forcing," *Northwestern Univ Law Rev.*, Mar./Apr. 1960, 55, 62-95.

G. A. Hay, "An Economic Analysis of Vertical Integration," *Ind. Org. Rev.*, 1973, *1*, 188-98.

L. W. McKenzie, "Ideal Output and the Interdependence of Firms," *Econ. J.*, Dec. 1951, *61*, 785-803.

M. K. Perry, "The Theory of Vertical Integration by Imperfectly Competitive Firms," memo. no. 197, Stanford Univ. Center Res. Econ. Growth, 1975.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ*, Apr. 1973, *81*, 442-49.

G. J. Stigler, "A Note on Block Booking,"

in Philip B. Kurland, ed., *The Supreme Court Review*, Chicago 1963.

J. M. Vernon and D. A. Graham, "Profitability of Monopolization by Vertical Integration," *J. Polit. Econ.*, Oct. 1971, *79*, 924-25.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, July/Aug. 1974, *82*, 783-802.

O. E. Williamson, "Economies as an Antitrust Defense: The Welfare Tradeoffs," *Amer. Econ. Rev.*, Mar. 1968, *58*, 18-36.

———, "The Vertical Integration of Production: Market Failure Considerations," *Amer. Econ. Rev Proc.*, May 1971, *61*, 112-23.

Fortner Enterprise, Inc. v. United States Steel Corp., 394 U.S. 495 (1969).

Northern Pacific Railway Co. v. United States, 356 U.S. 1 (1958).

Times-Picayune Publishing Co. v. United States, 345 U.S. (1953).

United States v Lowe's, Inc., 371 U.S. 38 (1962).

United States Steel Corporation v. Fortner Enterprises, Inc., 45 L. W. 4171 (1977).

Dynamic Models of Portfolio Behavior: More on Pitfalls in Financial Model Building

By DOUGLAS D. PURVIS*

In an important article in this *Review*, William Brainard and James Tobin have emphasized the role played by the wealth constraint in systems of asset demand equations. The wealth constraint gives rise to consistency conditions which must be satisfied by the demand functions when such a system is specified and estimated. As Brainard and Tobin caution, care must be taken to ensure that unrealistic coefficients are not inadvertently imposed on omitted equations by failure to recognize the consistency conditions.[1] Noting that the wealth constraint applies out of, as well as in, portfolio equilibrium, Brainard and Tobin focus attention on systems in which actual and desired stocks of assets differ. They specify a multivariate stock adjustment model wherein the desired change in holdings of any asset depends in general upon all asset stock disequilibria; the existence of such stock disequilibria can be implicitly rationalized on the basis of costs of adjustment which impinge on the rate of change of at least some assets. In this framework they show that the stock adjustment coefficients must also satisfy certain consistency conditions to ensure that the wealth constraint is satisfied.

An important feature of their analysis is that the total change in wealth (savings plus capital gains) is treated as exogenous to the financial sector, and the asset flow demands described above are *conditional* upon the exogenously given change in wealth. This strategy of separating the portfolio balance decision from the consumption-saving decision is one that Tobin has explicitly used and justified in his 1969 article (especially pp. 15 16), and is one that has been widely and effectively used in modern macroeconometric models.

The central argument of the present paper is that this separation of flow-allocation and stock-allocation decisions is not legitimate in the presence of adjustment costs attached to changing the level of individual asset holdings. The existence of adjustment costs means that there is no portfolio balance problem per se (in the sense of allocation of a given level of wealth), but rather a (longer run) problem of determining an optimal time path for each asset and for the level of consumption. Thus a natural extension of the Brainard-Tobin model is to treat saving and portfolio decisions in an integrated fashion.[2]

Note that the Brainard-Tobin model is perfectly consistent with *any* model of savings behavior and hence no logical con-

*Queen's University, and Cowles Foundation for Research in Economics, Yale University. I am grateful to Adrian Pagan, Gordon Sparks, and James Tobin for helpful discussions, and especially to Gary Smith who, as well as patiently discussing many of the issues, provided detailed comments on earlier drafts of this paper. This research was partially supported by a National Science Foundation grant to the Cowles Foundation and by a Canada Council grant to the author Remaining mistakes and opinions are my own.

[1]This also has implications for the common practice in macro-economic models of leaving the bond market as implicit. Care must be taken to ensure that silly behavior is not inadvertently attributed to bondholders. William Silber, Tobin, and Alan Blinder and Robert Solow have initiated research which "reintroduces" the bond market into macroeconomic models.

[2]It appears to be a fairly general result that the existence of adjustment costs leads to integrated behavior. M. Ishaq Nadiri and Sherwin Rosen have established a similar result for the theory of the firm, and Robin Mukherjee and Edward Zabel have recently shown that the "separation theorem" prominent in the finance literature on the mean-variance approach to optimal consumption-portfolio behavior fails to hold when transactions costs are introduced. In my 1975 paper (Appendix), I have argued that the integration of saving and portfolio balance decisions also applies in continuous-time models, even though such models are characterized by separate stock and flow budget constraints.

tradiction with an integrated approach is necessarily involved. However, by the very fact that their analysis is consistent with *any* model of saving behavior, we observe that it fails to incorporate information contained in the implicit assumptions about adjustment costs, and this information plays a central role in the integrated model presented in detail in Section I. Then, in Section II the Brainard-Tobin model is reexamined in light of the results of Section I Section III offers some concluding comments on how the integrated model presents a framework in which one can discriminate between monetarist and Keynesian views about the transmission of monetary policy.

I. An Integrated Model of Consumption and Asset Accumulation

As noted above, the wealth constraint plays a central role in the Brainard-Tobin analysis. However, it is worth noting that in the discrete time framework generally used there is no wealth constraint per se, but rather one single budget constraint relating assets held at the beginning of the period plus income received during the period to assets held at the end of the period plus consumption during the period. The budget constraint is given by

$$(1) \qquad \sum_r y_r + c = \sum_r y_r(-1) + X_0$$

where $y_r(-1)$ and y_r are, respectively, beginning and end of period holdings of the rth asset, c is consumption, and X_0 is income.[3]

To specify asset demands subject to a wealth constraint in such a model is to assume that the household makes its saving decision thereby determining end of period wealth ($X_m = \Sigma_r y_r(-1) + X_0 - c$), and *independently* chooses the desired allocation of that wealth amongst the alternative assets. This specification is reasonable when there are no adjustment costs pertaining to the reallocation of a given level of wealth. For then the household's initial position is fully described by the level of income X_0 and the *total* of assets inherited from the previous period $\Sigma y_r(-1)$. Saving and portfolio balance decisions may be influenced by common variables, but in this case both will be independent of the composition of inherited wealth.[4] However, when such adjustment costs are present, knowledge of the household's initial position requires knowledge of its income and the values of individual assets $y_r(-1)$ for *each r*. The rational household would, in such circumstances, formulate consumption and asset flow demands dependent upon income, current holdings of individual assets, and long-run asset considerations. For simplicity I assume that the latter are captured by steady-state demands which are independent of initial asset holdings. It is assumed that the desired asset positions y_r^* depend only on the expected values of income and rates of return $X_i^e (i \neq m)$, which in turn, assuming static expectations, equal actual $X_i (i \neq m)$. Throughout, I abstract from uncertainty so that all changes in wealth are planned. Consumption and the asset flow demands are given, respectively, as[5]

$$(2) \qquad c = \sum_i b_i X_i + \sum_s e_s y_s(-1)$$

[3]Throughout I will use the following conventions: X_m is total wealth; X_0 is income; subscripts (i, j) will be used to refer to explanatory variables (X) and unless otherwise noted will take on values from 0 to $m - 1$ in summation; and subscripts (r, s) will refer to assets (y) and take on values 1 to n in summation.

[4]In the absence of adjustment costs, asset demands would be short-run in nature (contingent upon total wealth) and would also be *achieved* in the short run. This is what Lewis Johnson has called a portfolio balance model as opposed to the portfolio adjustment model used in the present paper.

[5]As the referee has pointed out, this and the Brainard-Tobin models are versions of what Duncan Foley has called an end-of-period equilibrium model. Note however that Foley's discussion is somewhat misleading the end-of-period specification does have an equilibrium This can be seen by observing that $K^D(0)$ in his first equation on p 312 is the short-run demand for capital while in equation (4a), where it appears multiplied by k, $K^D(0)$ is the long-run demand, in both, the short-run excess demand disappears as the time period goes to zero.

$$(3)\quad \Delta y_r = \sum_s \gamma_{rs}[y_s^* - y_s(-1)] \qquad \text{for } r = 1,\ldots,n$$

The target values y_s^* are *long-run* steady-state demands,

$$(4)\quad y_s^* = \sum_i \beta_{si} X_i \qquad \text{for } s = 1,\ldots,n$$

Note that by assumption current wealth (X_m) does *not* enter the long-run asset demand functions (4). However, the n long-run asset demand functions together do imply a target value for total wealth X_m^*, and long-run equilibrium, given X_m^*, will be independent of whether the composition of wealth matters in the short run. Substituting from (4) into (3) yields the n reduced form asset-demand equations for estimation

$$(5)\quad \Delta y_r = \sum_i \alpha_{ri} X_i - \sum_s \gamma_{rs} y_s(-1) \qquad \text{for } r = 1,\ldots,n$$

where the impact multipliers given by $\alpha_{ri} = \Sigma_s \gamma_{rs}\beta_{si}$ denote the total impact effect of a change in the ith explanatory variable on the flow demand for the rth asset.

The $(n + 1)$ equations (2) and (5) determine the $(n + 1)$ endogenous variables $\Delta y_r(r = 1,\ldots,n)$ and c. There are n predetermined variables $y_s(-1)$, and m exogenous variables $X_i(i = 0,\ldots,m - 1)$. Total wealth does not appear explicitly since it is the composition of wealth which is important in determining the optimum time paths. There are only n independent endogenous variables since consumption is linked to the asset flow demands by the budget constraint given by (1) and rewritten below in flow terms as

$$(6)\quad X_0 = c + \sum_r \Delta y_r$$

This introduces a linear dependency among the $(n + 1)$ equations which is reflected in the following $(m + n)$ consistency conditions:

$$(7a)\quad b_0 = \sum_r \alpha_{r0} = 1 \quad (\text{since } X_0 \text{ is income})$$

$$(7b)\quad b_i + \sum_r \alpha_{ri} = 0 \qquad \text{for } i = 1,\ldots,m - 1$$

$$(7c)\quad e_s - \sum_r \gamma_{rs} = 0 \qquad \text{for } s = 1,\ldots,n$$

These consistency conditions ensure that the budget constraint is always satisfied: (7a) simply reflects the fact that income gets allocated to either consumption or asset accumulation; (7b) and (7c) show that changes in nonincome explanatory variables (rates of return or initial stocks) cause a change in the allocation of income, consistent with the total ($c + \Sigma\,\Delta y_r$) being constant.

Of particular interest is (7c) which states that the column sums of the spillover matrix $[\gamma_{rs}]$ must vary to reflect any variation in the coefficients of the lagged stocks in the consumption equation. Variations in these latter coefficients indicate how the composition of assets influences consumption; variations in the column sums of the spillover matrix are the "mirror images" which indicate how the composition of assets influences saving. Saving, found by summing over the asset flow demand functions, using (7c) is given by

$$(8)\qquad \Delta X_m = \sum_s e_s[y_s^* - y_s(-1)]$$

The special case where only total wealth influences consumption, which obtains in the absence of adjustment costs or possibly under restrictive assumptions about the underlying structure of such costs or behavior, arises when $e_s = \bar{e}$ for all s. From (7c) we can see this means that the column sums $\Sigma_r \gamma_{rs}$ are constant at e for all s, the multivariate analogue of a constant speed of adjustment.[6]

The model outlined above is similar in structure to that presented by Brainard and

[6]Saving, in this case, can be represented as the adjustment of actual towards desired wealth,

$$\Delta X_m = \bar{e}[X_m^* - X_m(-1)] \qquad \text{where } X_m^* = \Sigma\, y_s^*$$

Tobin, but, as a result of explicit consideration of the existence of adjustment costs, differs in several respects, especially with regard to the role played by the wealth constraint. However, it is still true that at any point in time, the sum of asset holdings must equal the value of financial wealth. That fact is incontrovertible; only the experiment giving rise to "consistency conditions" is changed since the change in wealth during any period is now treated endogenously. Furthermore, the basic Brainard-Tobin message remains intact; the potential pitfalls awaiting the researcher who does not take full account of the implications of the budget constraint are severe.

II. The Brainard-Tobin Pitfalls Model

In terms of the notation set out above, the Brainard-Tobin model can be illustrated in terms of the n asset equations given by

$$(9) \quad \Delta y_r = \sum_s \hat{\gamma}_{rs}[y_s^* - y_s(-1)] + \delta_r \Delta X_m \qquad \text{for } r = 1, \ldots, n$$

where the asset targets y_s^* are given by

$$(10) \quad y_s^* = \sum_{i=0}^{m} \beta_{si} X_i \qquad \text{for } s = 1, \ldots, n$$

and where $X_m = \Sigma y_s = \Sigma y_s^*$. There are three basic differences between the model described by (9) and (10) and the model outlined in Section I above. These are: the inclusion of the term $\delta_r \Delta X_m$ on the right-hand side of (9); the lack of an equation for saving; and the inclusion of end of period wealth X_m as an explanatory variable in (and constraint on) the asset target demand functions given by (10).

The first point has been adequately discussed by Gary Smith; we only need to note that since the $\delta_r \Delta X_m$ term is completely redundant we can, without loss of generality, drop it. As Smith notes, this is equivalent to using the substitution $\Delta X_m = \Sigma[y_s^* - y_s(-1)]$ to derive new stock adjustment coefficients $\gamma_{rs} = \hat{\gamma}_{rs} + \delta_r$. This is only one of many possible substitutions, all of which are equivalent, as shown by Smith on page 512.[7]

The second point is more important for the comparative interpretation of the two models. By analyzing the consumption-saving decision separately, the Brainard-Tobin model treats the change in wealth, that is, the sum over r of the left-hand side of equation (9), as exogenous to the financial sector. This gives rise to the consistency conditions shown by (11):

(11a)

$$\sum_r \alpha_{ri} = 0 \qquad \text{for } i = 0, 1, \ldots, m-1$$

(11b)

$$\sum_r \alpha_{rm} = 1$$

(11c)

$$\sum_r \gamma_{rs} = 1 \qquad \text{for } s = 1, \ldots, n$$

where the alphas are impact multipliers as described above. Conditions (11) reflect both the fact that asset demands must satisfy the wealth constraint *and* the treatment of wealth as exogenous. Thus (11a) shows that changes in nonwealth explanatory variables lead only to a reshuffling of assets since wealth is explicitly held constant; similarly (11b) shows that an exogenous increase in wealth must induce an equivalent increase in asset holdings. Condition (11c) implies that any reshuffling of initial assets $y_s(-1)$, holding $X_m(-1)$ constant, does not *necessarily* influence current wealth X_m. This of course must be the case since X_m is exogenous. In contrast to the role played by the column sums in the integrated model of Section I (i.e., condition

[7] Brainard-Tobin substitute for the excess demand of equities They originally introduced the redundancy shown in equation (9) to allow for differential effects of alternative sources of wealth changes (i.e., planned savings vs. capital gains) and hence *two* terms were added to the stock adjustment terms. Since equations such as (9) don't make this distinction, the additional term is completely redundant; as Smith's clarification shows, the confusions in the literature that he addresed were completely unnecessary.

(7c)), this condition tells us nothing about the influence of the composition of wealth on consumption.

It is important to emphasize the different nature of the "flow-asset demand" equations (3) of the integrated model compared to the "contingent-asset-flow" equations (9). If the latter (assuming $\delta_r = 0$) were interpreted as asset-flow demands like the former, and hence their sum were used to represent the *planned* change in wealth, then conditions (11) would imply that saving not only be independent of the composition of wealth (11c), but also of income and rates of return (11a). This, of course, is *not* the way these equations were meant to be interpreted. The point is that, due to the exogeneity of wealth, conditions (11) are only sufficient conditions; if a saving relation explaining ΔX_m is added to the model depicted by (9) and (10), then the sufficient conditions (11) can be replaced by a (possibly less restrictive) set of necessary conditions which will depend upon the saving behavior postulated, as illustrated in the integrated model set out above.[8]

The second point also arises when it comes to implementing the Brainard-Tobin model. There will be econometric problems involved in using wealth as an explanatory variable and hence treating it as predetermined, if it is in fact systematically related to the other explanatory variables. In order to deal with this, one would likely resort to limited information techniques, such as instrumental variables; the natural instruments to choose would be the variables expected to influence saving. Hence the econometric problem is really a reflection of the specification; the integrated model which explicitly includes the consumption-saving choice would lend itself to full-information estimation procedures.

The third basic difference between the models is that Brainard-Tobin constrain the target values by end of period wealth. Mark Ladenson presents this as "... a sort of rational desires hypothesis [since it] constrains the *desired* or equilibrium values of the financial assets and liabilities to obey the balance sheet identity" (p. 179). Such a statement is misleading since the target demand y_s^* is not in general achieved at the end of the period and hence is not an "effective demand"—there is no reason for it to be constrained by end of period wealth. Obviously what must be constrained are the actual quantities y_s. Of course, the model is all conditional upon X_m, and the y_s^* are the quantities that ultimately would be held if the values of X_m were constant for a long enough time. But, there is no reason for (expected) X_m to remain constant, and the decision rules (9) are very myopic in that they don't take into account possible changes in the size of total wealth when planning the adjustment paths of individual assets. In the integrated model, the desired paths for assets do incorporate the accumulation decision, and total wealth at the end of period is as much a result of as it is a constraint on asset demands.[9]

Hence the Brainard-Tobin pitfalls model is formally consistent with the integrated model studied in Section I, when combined with a consumption-savings relationship such as (2), the Brainard-Tobin model will in principle give rise to exactly the same short- and long-run behavior as the integrated model.[10] However, the research

[8] Note that this issue arises independently of whether saving depends on the composition of wealth. If one is working with only a model of the financial sector, conditions (11) are appropriate. If, however, one is working with a complete macro model including a saving function, then to continue to impose conditions (11) will in general result in ignoring information present elsewhere in the system.

[9] Dropping X_m as an explanatory variable for y_s^* also eliminates the redundancy in equation (9). However we continue to omit the $\delta_r \Delta X_m$ terms since ΔX_m is endogenous. Constraining the y_s^* by X_m is the feature of the Brainard-Tobin model which renders the univariate stock adjustment model ($\gamma_{rs} = 0$ for $r \neq s$) inconsistent. In the integrated model of Section I, univariate adjustment is possible ($\gamma_{rr} = e_r$; $\gamma_{rs} = 0$, $r \neq s$) as is the special case of constant speed of adjustment ($\gamma_{rr} = \bar{e}$)

[10] In practice, the results may differ since the "familiar" pitfalls model would maintain a deterministic relationship between $\Sigma\, \Delta y_r$ and ΔX_m but, presumably, a stochastic relationship between ΔX_m and $X_0 - c$. The integrated model would exclude the concept ΔX_m and posit a deterministic relationship between $\Sigma\, \Delta y_r$ and $X_0 - c$.

strategies underlying the two approaches are very different, and care must be taken to interpret the results of the models accordingly.

III. Concluding Comments

The portfolio balance models common in the literature have been based on a dichotomized approach to household decisions; as a result the coefficients of the asset demands were not only constrained to satisfy the wealth constraint but also to be consistent with the exogenous wealth assumption. I have presented an integrated approach to household behavior and developed a model which was more general but no more complicated than the portfolio balance model; in fact, the latter model is in some sense a special case of the integrated model.

These results are important for the analysis of the linkage between the monetary and real sectors of the economy. Tobin has used the dichotomized model to provide an insightful analysis of how changes in relative asset supplies alter the structure of equilibrium rates of return (especially the supply price of capital) which in turn induce expenditure flows. Thus monetary policy operates in the usual Keynesian fashion by influencing interest rates; since total wealth is unchanged, no direct effect on expenditure is postulated. Monetarists, on the other hand, suggest that changes in relative asset supplies may elicit direct expenditure effects so that an open market purchase, for example, may affect expenditure other than via interest rates. That is, the monetarist view maintains that there will be a "real balance effect" even in the presence of a multiasset portfolio. In terms of my model the monetarist view of the transmission process is that the coefficient of money in the consumption equation would exceed that on other assets, a possibility generally ruled out in the dichotomized Keynesian model.[11] The integrated model presents a framework in which the monetarist view and the Keynesian view arise as special cases, but the special cases can be tested for rather than imposed a priori.

In the integrated model, consumption would in general be affected by the process of the adjustment of actual toward desired asset holdings; a change in asset stocks may well alter current consumption in a manner independent of changes in current or desired wealth (the latter arising out of interest rate changes) by altering the "speed" with which current adjusts to desired wealth. The monetarist view of the transmission process suggests, in terms of my model, that a substitution of money for any other asset increases consumption, that is, it slows the adjustment of actual towards desired wealth. This "liquidity" effect is a short-run dynamic effect since in my model no effect on long-run wealth and consumption is postulated. But in the short run there is a close relationship between money and expenditure, and in this sense, velocity is relatively constant in the short run.[12]

A more general result from the considerations in this section is that consistency conditions derived from examining the portfolio balance decision can be treated only as "sufficient" conditions. Once systematic relationships between the explanatory variables are introduced in other equations—for example, the savings decision—then the sufficient conditions can be replaced by a less restrictive set of "necessary" conditions. Such conditions allow not only for more general models of consumption behavior as emphasized above, but also give rise to an estimation procedure in which asset substitution relationships may

[11] This is an empirical rather than a theoretical proposition since there is nothing in my model to distinguish money from any other asset. Milton Friedman has argued that the monetarist view of the transmission process does not confine the substitution effects to a limited range of financial assets, but supposes that individuals seeking "to dispose of what they regard as their excess money balances ... will try to pay out a larger sum for the purchase of securities, goods and services, for the repayment of debts, and as gifts than they are receiving from corresponding sources" (p. 910).

[12] See also Friedman's comment, p. 316, in Jerome Stein's volume, on the relative stability of stock-flow and flow-flow relationships.

appear very different from those resulting in the standard approach since the constraints on such relationships are now relaxed somewhat (compare (7a) and (7b) with (11a) and (11b)).

REFERENCES

A. Blinder and R. Solow, Does Fiscal Policy Matter?," *J. Publ. Econ.*, Aug. 1973, *1*, 319-37.

W. C. Brainard and J. Tobin, "Pitfalls in Financial Model Building," *Amer. Econ Rev. Proc.*, May 1968, *58*, 99-122.

D. Foley, "On Two Specifications of Asset Equilibrium in Macroeconomic Models," *J. Polit. Econ.*, Apr. 1975, *83*, 303-24.

M. Friedman, "Comments on the Critics," *J Polit. Econ.*, Sept./Oct 1972, *80*, 906-50.

L. Johnson, "Inflationary Expectations and Monetary Equilibrium," *Amer Econ Rev.*, June 1976, *66*, 395-400.

M. Ladenson, "Pitfalls in Financial Model Building: Some Extensions," *Amer Econ Rev.*, Mar. 1971, *61*, 179-86.

R. Mukherjee and E. Zabel, "Consumption and Portfolio Choices with Transactions Costs," in Michael Balch et al., eds., *Essays on Economic Behavior Under Uncertainty*, Amsterdam 1975, ch. 6.

M. J. Nadiri and S. Rosen, "Interrelated Factor Demand Functions," *Amer. Econ. Rev.*, Sept. 1969, *59*, 457-71.

D. D. Purvis, "Portfolio and Consumption Decisions: Towards a Model of the Transmission Process," paper presented to Reserve Bank of Australia Conference in Monetary Economics, Sydney, July 1975

W. Silber, "Fiscal Policy in an IS-LM Model: A Correction," *J. Money, Credit, Banking*, Nov. 1970, *2*, 461-72.

G. Smith, "Pitfalls in Financial Model Building: A Clarification," *Amer. Econ Rev.*, June 1975, *65*, 50-516.

Jerome Stein, *Monetarism*, Amsterdam 1976.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15-29.

Dynamic Models of Portfolio Behavior: Comment on Purvis

By GARY SMITH*

Douglas Purvis' discussion of an integrated approach to consumption and portfolio decisions is an attractive extension of the "pitfalls" framework advocated by William Brainard and James Tobin. The pitfalls model is concerned with the portfolio allocation of a level of wealth which is predetermined by beginning of period asset holdings and current period saving and capital gains. One of the innovative features of this model is the inclusion of all asset yields and lagged asset holdings as explanatory variables in the asset demand equations. Purvis supplements the Brainard-Tobin asset demands with a consumption-saving relationship that includes a similar list of explanatory variables and reinterprets this system as a model of integrated rather than sequential decision making. Despite his observation that, "when combined with a consumption-savings relationship such as (2), the Brainard-Tobin model will in principle give rise to exactly the same short- and long-run behavior as the integrated model" (p. 407), most of Purvis' discussion is concerned with alleged dissimilarities between the two approaches. This is apparently due to his implicit coupling of a simple consumption function and sequential decision making. In particular most of his comments on the Brainard-Tobin approach are actually concerned with whether or not lagged asset holdings should be included in a consumption function. This is rather unfair to Brainard and Tobin since there is no consumption function in the pitfalls model, and the two issues are really conceptually distinct. An integrated approach does not preclude, and a sequential approach does not require, a simple consumption function. The spirit of Brainard and Tobin's work is in fact that the inherited composition of wealth is very important to consumption, but consumption decisions precede asset demand decisions. The substance of their sequential approach is not that the composition of wealth is unimportant to consumption but rather that there are some variables which influence consumption and yet do not separately affect asset demands; only the net amount of saving motivated by these influences is important. In this paper I have consequently tried to separate these two issues: the use of an integrated or sequential framework and the imposition of parametric assumptions.

One of the reasons for the merging of these two issues in Purvis' discussion is that he uses a deterministic scenario which makes the distinction between integrated and sequential decisions unimportant. In Purvis' integrated model, consumption and asset demands are constrained by lagged asset holdings plus income. In the relevant sequential interpretation of this model, consumption is first determined, setting the amount of saving and the level of end of period wealth. Asset demands are then decided upon, subject to the budget constraint that they sum to the predetermined end of period wealth. Thus the integrated asset demands include income as an explanatory variable while the sequential asset demands instead include end of period wealth. In a deterministic world there are no substantive differences between these approaches as long as income and wealth are related through a consumption-saving equation. This equivalence breaks down if the marginal propensity to save out of income is zero (since wealth is then no longer related to income) or if there is an unobserved disturbance term in the consumption

*Yale University. Note that equations numbered (1) through (11) are in Purvis' paper. My equations are numbered in the same sequence.

equation. With a stochastic error term, predictions will diverge and the use of observed wealth in the asset demand equations will introduce biases into the estimation procedure unless the decision making is truly recursive. If there is a simultaneous equations problem, then the use of the consumption explanatory variables as instruments for wealth will restore the equivalence of the sequential and integrated models.

In practice, the attractiveness of a sequential approach will depend upon the sector being studied. A hierarchical approach in which certain decisions are prior claims which constrain other decisions is often used with considerable success despite its weak theoretical underpinnings. Some examples are household saving preceding the allocation of wealth, deposits in financial institutions constraining their asset acquisition, corporate physical investment preceding financing, and corporate investment and long-term financing preceding short-term asset management.

Although it need not, a sequential approach is in practice often used to motivate the imposition of parametric restrictions. Some extreme examples are the assumptions that corporate investment depends only upon a comparison of the anticipated profit rate with some hurdle rate; that the supply of labor depends only upon the nominal or real wage rate; and that consumption depends only upon disposable income. This practice may underlie Purvis' implicit assumption that his complicated consumption function would not be used in a sequential model. I am instead stressing the points that a sequential approach does not require a simple consumption function and that a comparison of consumption functions does not provide a test of a sequential approach.

The actual Brainard-Tobin asset demands involve somewhat different parametric assumptions in that both income and wealth are included as explanatory variables. In Purvis' deterministic model, this is redundant since the saving relation makes wealth a linear function of income and the other explanatory variables. In a stochastic world, wealth would pick up unexplained saving. In practice the substantive assumption actually embodied in the Brainard-Tobin sequential approach is that a number of explanatory variables in the consumption function do not separately appear in the asset equations but instead influence asset holdings only through wealth. Thus wealth appears in the asset demand equations because a number of other variables have been omitted from these equations, and not because lagged asset holdings have been omitted from the consumption function.

In order to separate in the present paper the selection of explanatory variables from the adoption of a sequential or integrated approach, I will first discuss the sequential vs. integrated argument using the Purvis model to more concretely illustrate the issues involved. I will then compare the Purvis model to the Brainard-Tobin asset demand equations and discuss the implications of the differences in explanatory variables.

I. The Relationship Between Integrated and Sequential Approaches

In order to analyze both deterministic and stochastic frameworks, I will add a stochastic disturbance term ϵ_0 to Purvis' consumption function (2) and stochastic terms ϵ_r to his asset flow demands (5). The presence of these terms introduces an additional adding up restriction to his equations (7a)-(7c):

$$\text{(7d)} \qquad \epsilon_0 + \sum_r \epsilon_r = 0$$

In order to reinterpret his integrated model as a sequential process, the consumption function (2) can be substituted into the budget constraint (6) to yield a relationship between income X_0 and end of period wealth $X_M = \Sigma_{s=1}^{m} y_s$ as long as the marginal propensity to save out of income $(1 - b_0)$ is not zero,

$$(12)\quad X_0 = \left(X_M + \sum_{i=1}^{m-1} b_i X_i + \sum_{s=1}^{n} (e_s - 1) y_s(-1) + \epsilon_0\right)/(1 - b_0)$$

The substitution of this relation into the asset demands (5) yields the sequential asset demands which depend upon wealth rather than income,

$$(13)\quad \Delta y_r = \frac{\alpha_{r0}}{1 - b_0} X_M + \sum_{i=1}^{m-1}\left(\alpha_{ri} + \frac{\alpha_{r0} b_i}{1 - b_0}\right) X_i - \sum_{s=1}^{n}\left(\gamma_{rs} + \frac{\alpha_{r0}(1 - e_s)}{1 - b_0}\right) y_s(-1) + \left(e_r + \frac{\alpha_{r0}}{1 - b_0}\epsilon_0\right)$$

$$= \frac{\alpha_{r0}}{\sum_r \alpha_{r0}} X_M + \sum_{i=1}^{m-1}\left(\alpha_{ri} + \frac{\left(\alpha_{r0}\sum_r \alpha_{ri}\right)}{\sum_r \alpha_{r0}}\right) X_i - \sum_{s=1}^{n}\left(\gamma_{rs} + \frac{\alpha_{r0}\left(1 - \sum_r \gamma_{rs}\right)}{\sum_r \alpha_{r0}}\right) y_s(-1) + \left(\epsilon_r - \frac{\alpha_{r0}}{\sum_r \alpha_{r0}} \Sigma \epsilon_r\right)$$

$$= \sum_{i=1}^{\bar{M}} \bar{\alpha}_{ri} X_i - \sum_{s=1}^{n} \bar{\gamma}_{rs} y_s(-1) + \bar{\epsilon}_r$$

The substitution of the definitions of the parameters contained in (13) into Purvis' adding up restrictions (7) for the integrated model yields the adding up restrictions for the sequential model,

$$(14a)\quad \sum_r \bar{\alpha}_{ri} = 0, \qquad i = 1, \ldots, m - 1$$

$$(14b)\quad \sum_r \bar{\alpha}_{rm} = 1$$

$$(14c)\quad \sum_r \bar{\gamma}_{rs} = 1, \qquad s = 1, \ldots, n$$

$$(14d)\quad \sum_r \bar{\epsilon}_r = 0$$

In a deterministic world where the $\epsilon_r = 0$, the $m + n$ explanatory variables in the sequential asset demands (13) are a linear transformation of the $m + n$ explanatory variables in the integrated demands (5), and the two approaches are fully equivalent. Notice that in contrast to Purvis, I have placed bars over the parameters of the sequential asset demands (13) to distinguish them from the integrated parameters in (5) because, although linearly related, they are not identical. The use of the same notation might give the misleading impression that the demand equations (5) and (13) and the corresponding consistency conditions (7) and (14) are competing alternatives rather than equivalent representations.

This equivalence breaks down in a stochastic world with unobserved disturbances, $\epsilon_r \neq 0$, since there will no longer be an exact linear relationship between explanatory variables for the integrated and sequential models. Even if one knew the true values of the parameters (or any common set for that matter), predictions would diverge since a unitary increase in the consumption disturbance term will not affect the integrated asset demand forecasts (5) but will lower X_m by a unit (see (12)) and hence lower the sequential asset predictions (13) by $\alpha_{r0}/(1 - b_0)$. The actual change in asset holdings will depend upon the correlation between ϵ_0 and ϵ_r (which is typically negative from (7d)). For any common set of parameters, the difference between the mean squared forecast errors for the sequential and integrated models is

$$\frac{\alpha_{r0}}{1 - b_0} E(\epsilon_0^2) + 2\frac{\alpha_{r0}}{1 - b_0} E(\epsilon_r \epsilon_0)$$

The integrated model will consequently be more accurate if and only if

$$(15) \qquad 2\frac{E(\epsilon_r\epsilon_0)}{E(\epsilon_0^2)} > -\frac{\alpha_{r0}}{1-b_0}$$

In practice the parameters will probably be estimated, and one must then confront the fact that wealth is likely to be correlated with the disturbance terms in the sequential asset demand equations, since

$$E(X_m\bar{\epsilon}_r) = -E(\epsilon_0\epsilon_r) - \frac{\alpha_{r0}}{1-b_0}E(\epsilon_0^2)$$

is not equal to zero unless

$$-\frac{E(\epsilon_0\epsilon_r)}{E(\epsilon_0^2)} = \frac{\alpha_{r0}}{1-b_0}$$

The left-hand side of this condition[1] describes the change in Δy_r associated with a unit increase in X_m that is due to a fall in ϵ_0. The right-hand side is the increase in Δy_r associated with a unit increase in X_m due to an increase in X_0. Thus wealth will be correlated with the disturbance terms in the asset demand equations, unless the model is actually sequential in the specific sense that the effect of a change in wealth on asset demands is independent of whether the change in wealth is the result of a change in income or in the consumption disturbance term.[2] The earlier prediction criteria (15) can now be interpreted as stating that (for a common set of parameter estimates) the sequential model will have smaller forecast errors if it is more accurate to assume that a change in the consumption disturbance term has the same effect as a change in income on asset demands than to assume that it has no effect.

An obvious response to the simultaneity problem is to use X_i $(i = 0,\ldots,m-1)$ and $y_s(-1)$ $(s = 1,\ldots,n)$ as instruments for wealth X_m. The sequential asset demand equations will then include $m + n$ independent linear combinations of the $m + n$ integrated explanatory variables, rendering the two sets of parameter estimates fully equivalent. The only difference will be the superficial one that one set of equations will be in the integrated form (2) and (5) with its parameters subject to the adding up restrictions (7) while the other set will be in the sequential form (2) and (13) with its parameters subject to (14). Either set of parameter estimates could of course be directly derived from the other set. The forecasts will also coincide if X_m is replaced by its instrumental variables estimate. If the observed values of X_m are instead used, then the forecasts will diverge with the criteria (15) again determining the more accurate approach. It is also worth mentioning that wealth is not the only explanatory variable which may not be predetermined. Asset prices or yields, the prices of consumption goods and, more relevantly, income could all be plausibly considered sources of simultaneity problems.

II. The Pitfalls Model

The Brainard-Tobin asset demands discussed by Purvis (from (9) and (10)) modify this analysis somewhat since both income and wealth are included as explanatory variables,

$$(16) \quad \Delta y_r = \sum_{i=0}^{m} \bar{\alpha}_{rm} X_i - \sum_{s=1}^{n} \bar{\gamma}_{rs} y_s(-1) + \bar{\epsilon}_r$$

The adding up restrictions are now[3]

[1] An alternative interpretation is that the system must be recursive in that the error term ϵ_0 in the consumption function is uncorrelated with the error terms $\epsilon_r + \alpha_{r0}\epsilon_0/(1-b_0)$ in the sequential asset demands (13)

[2] Purvis misstates the simultaneity issue as, "There will be econometric problems involved in using wealth as an explanatory variable and hence treating it as predetermined if it is in fact systematically related to the other explanatory variables" (p. 407) Simultaneity does not hinge upon the correlation between wealth and the other explanatory variables, but rather upon the correlation between wealth and the disturbance term.

[3] Purvis argues that the Brainard-Tobin adding up restrictions "reflect . the treatment of wealth as exogenous, . " and that, "Condition [(17c)] implies that any reshuffling of initial assets $y_r(-1)$, holding $X_m(-1)$ constant does not *necessarily* influence current wealth X_m" (p 406). The model and the restrictions in fact contain no information about the determination of wealth. Conditions (17a), (17c), and (17d) do reflect a mental *ceteris paribus* experiment in which wealth is held constant while another ex-

$$\text{(17a)} \quad \sum_r \bar{\alpha}_{ri} = 0, \qquad i = 0, \ldots, m-1$$

$$\text{(17b)} \quad \sum_r \bar{\alpha}_{rm} = 1$$

$$\text{(17c)} \quad \sum_r \bar{\gamma}_{rs} = 1, \qquad s = 1, \ldots, n$$

$$\text{(17d)} \quad \sum_r \bar{\epsilon}_r = 0$$

I have again labelled the parameters with bars, both for notational simplicity and because the sequential version (13) of Purvis' model can be interpreted as the special case of the Brainard-Tobin model (16) in which $\bar{\alpha}_{r0} = 0$. This implies that since Purvis' integrated model, (2) and (5), is of comparable generality to the sequential version, (2) and (13), he is incorrect in arguing that the Brainard-Tobin model "is in some sense a special case of the integrated model" (p. 408), and that with an integrated approach the adding up restrictions "are now relaxed somewhat (compare (7a) and (7b) with ... [(17a)] and ... [(17b)])" (p. 409). The latter argument is also directly refuted by the observation that (7) is a linear transformation of the sequential adding up restrictions (14) which are the special case of (17) where $\bar{\alpha}_{r0} = 0$.[4]

An alternative interpretation is provided by using the implication of the consumption function (2) that

$$\text{(18)} \quad X_m = \sum_1^n y_s(-1) + X_0 - C = (1 - b_0)X_0 - \sum_1^{m-1} b_i X_i + \sum_1^n (1 - e_s)y_s(-1) - \epsilon_0$$

to eliminate X_m from (16):

$$\text{(19)} \quad \Delta y_r = (\bar{\alpha}_{r0} + \bar{\alpha}_{rm}(1 - b_0))X_0 + \sum_1^{m-1} (\bar{\alpha}_{ri} - \bar{\alpha}_{rm}b_i)X_i - \sum_1^n (\bar{\gamma}_{rs} - \bar{\alpha}_{rm}(1 - e_s))y_s(-1) + (\bar{\epsilon}_r - \bar{\alpha}_{rm}\epsilon_0)$$

This is identical to Purvis' integrated asset demands (5).

In a deterministic world the Brainard-Tobin model (16) and the Purvis model (19) are thus equivalent if the consumption function (2) is appropriate. This is because the consumption function (2) implies that X_m is redundant since it is linearly related to the remaining explanatory variables in the asset demand equations. As Purvis notes the adding up restrictions (17) are then sufficient but not necessary. This is analytically equivalent to the Ladenson-Clinton problem discussed by the author, though one could draw a distinction between linear dependencies arising from identities and those due to behavior and/or events. With the former, one can only use fanciful terms to describe behavior in the Wonderland situation where an explanatory variable is increased while it is also being held constant. In the latter case, however, it is entirely reasonable (and even interesting) to speculate on how agents might behave if for legal, policy, or behavioral reasons variables which had been tied together were to be set free. When engaging

planatory variable changes, but such experiments can be performed regardless of whether or not wealth is exogenous. For example, condition (17a) states that holding wealth as well as the other explanatory variables constant, an increase in X_i involves only a reshuffling of asset holdings. Purvis' condition (7b) on the other hand states that holding the explanatory variables other than wealth constant, an increase in X_i will increase net asset holdings by the amount that saving increases. If one were to perform comparable mental experiments, one would obtain comparable answers. For example, the asset demand effects of a change in X_i in the Brainard-Tobin model, taking into account the effect of X_i on wealth, also sum to $-b_i$.

[4]A tangential issue raised by Purvis is that the pitfalls framework is best suited for a static environment since the target asset holdings and adjustment behavior do not reflect the anticipated course of asset demands and resources. In a growth situation, the model is consequently always struggling to catch up to a moving target. It is unfortunately difficult to construct a simple and yet sophisticated description of this dynamic multivariate problem, although attempts have been made by Manuel Barbosa and Benjamin Friedman. Since Purvis' asset demands are less general than those of Brainard and Tobin, his contribution here is limited to his advocacy of the consumption function (2).

in such an exercise, one should assign parameter values which would yield consistent behavior if the variables were to move independently. The full set of adding up restrictions (17) would then be necessary as well as sufficient. In practice, the linear dependency in (16) will be broken if the parameters of the consumption function (2) change or if there is a stochastic error term in that equation. In a stochastic world, even with the common set of parameters described in (19), predictions will diverge[5] since an increase in the consumption disturbance term ϵ_0 will not affect the Purvis forecasts (19), but will lower wealth and hence the Brainard-Tobin asset forecasts (16) by $\overline{\alpha}_{rm}$. The actual mean change in Δy_r of $-\overline{\alpha}_{rm} + cov(\epsilon_0\epsilon_r)/var(\epsilon_0)$ is generally assumed negative.[6] In their predictions (and estimation) Brainard-Tobin implicity assume that $cov(\epsilon_0\bar{\epsilon}_r) = 0$ and assign a value $\overline{\alpha}_{rm}$ to the change in Δy_r. Purvis implicitly assumes that $cov(\epsilon_0\bar{\epsilon}_r)/var\,\epsilon_0 = \alpha_{rm}$ and predicts no change in Δy_r. Purvis will have a smaller mean squared error if and only if

$$\frac{cov(\epsilon_0\epsilon_r)}{var(\epsilon_0)} > \frac{\overline{\alpha}_{rm}}{2}$$

Thus, there is here a substantive difference in that Purvis neglects, while Brainard-Tobin do not, the effects on asset demands of influences which are omitted from the consumption function.

This brings up the more general point that Brainard and Tobin include wealth in the asset demands to catch the effects not only of ϵ_0, but also of any explanatory variables in the consumption function which are not otherwise included in the asset demand equations. The Brainard-Tobin asset demands are exceptionally rich and detailed, but they typically include only income, saving, capital gains, asset yields, and lagged asset stocks as explanatory variables, and neglect such obvious influences on saving as the composition of income, commodity prices, lagged stocks of durable commodities, and accustomed consumption or production levels. Instead they have adopted a sequential approach in which asset demands depend upon available saving, but not upon all of the factors which determine the level of saving. Thus the appearance of wealth in the asset demands does not reflect the simplifying assumption that asset stocks are unimportant to consumption but rather the simplifying assumption that many consumption influences are not separately important to portfolio decisions.

Although the Brainard-Tobin model reflects a simplification of asset demands rather than consumption, the presence of lagged asset stocks in the consumption function is interesting and deserves discussion.[7] Purvis specifically raises the issue of the monetarist transmission mechanism, although his mental experiment in which households find themselves with more money and fewer bonds reflects a benevolent mugger analogy rather than an open market operation in which households are persuaded to make a voluntary exchange.

The specific issue is the relative size of the coefficients e_s on the lagged asset stocks in the consumption function (2). If the coefficient for money is larger than that for bonds then a *ceteris paribus* swap of money for bonds will increase consumption. The coefficient e_s measures the reduction in saving occurring when the actual holdings of an asset increase relative to desired holdings. This coefficient $e_s = \Sigma_r\gamma_{rs}$ represents not only the own speed of asset adjustment (γ_{ss}), but also the induced changes in holdings of other assets (γ_{rs}, $r \neq s$). If these cross effects are negative then $e_s > \gamma_{ss}$, since saving must finance not only the acquisition of the own asset but of other assets as well. In the case of money one would expect the own effect to be large and the cross effects to be small but more likely negative than positive; other assets might be sold to obtain money but there is little reason to acquire other assets. With more

[5]If X_m itself is predicted from (2), then the asset demand forecasts will of course coincide

[6]In Purvis' framework, this assumption is that $cov(\epsilon_0\epsilon_r) < 0$

[7]Lawrence Klein; Don Patinkin; James Tobin; Harold Watts and Tobin; Arnold Zellner, David Huang, and L C Chau discuss the influence of the composition of asset holdings on consumption.

illiquid items, the own speed of adjustment is probably smaller but there may be significant positive cross effects; a hesitancy to move quickly into illiquid high transactions-cost items may motivate the temporary acquisition of liquid assets.

My own priors are that the e_s for liquid items are probably larger than for illiquid ones due to the extra transactions costs of using liquid assets as a temporary buffer for illiquid transactions. Within the liquid and illiquid and high and low transactions costs categories, the differences in e_s are probably negligible. Brainard and Tobin argued that model building often requires generality in order to avoid an inadvertantly implausible or even inconsistent specification. However, one should also seek to avoid the uselessness that follows from completely general models. Some subtle effects are surely so minor that they can be safely neglected.

REFERENCES

M. Barbosa, "Growth, Migration and the Balance of Payments in a Small Open Economy," unpublished doctoral dissertation, Yale Univ. 1977.

W. Brainard and J. Tobin, "Pitfalls in Financial Model Building," *Amer. Econ. Rev. Proc.*, May 1968, *58*, 99-122.

B. Friedman, "Financial Flow Variables and the Short-Run Determination of Long Term Interest Rates," unpublished paper, Harvard Univ. 1976.

Lawrence R. Klein, *Contributions of Survey Methods to Economics*, New York 1954.

Don Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965, Note M, 651-64.

D. Purvis, "Dynamic Models of Portfolio Behavior: More on Pitfalls in Financial Model Building," *Amer. Econ Rev.*, June 1978, *68*, 403-09.

G. Smith, "Pitfalls in Financial Model Building: A Clarification," *Amer. Econ. Rev.*, June 1975, *65*, 590-616.

J. Tobin, "Asset Holdings and Spending Decisions," *Amer Econ Rev. Proc.*, May 1952, *42*, 109 23.

H. W. Watts and J. Tobin, "Consumer Expenditures and the Capital Account," *Proc Conference Consumption Saving*, Vol. 2, Philadelphia 1960.

A. Zellner, D. S. Huang, and L. C. Chau, "Further Analysis of the Shortrun Consumption Function with Emphasis on the Role of Liquid Assets," *Econometrica*, July 1965, *33*, 571 81.

On Extortion: A Reply

By HAROLD DEMSFTZ*

In their criticism of my views on the extortion problem, George Daly and J. Fred Giertz confine their search for my opinion to a paragraph contained in a short comment I wrote (1971) on a paper by Donald Shoup. In that comment I referred the reader to a more extensive treatment of the subject (1972a). Had Daly and Giertz bothered to follow this lead they would have realized that their criticism completely misrepresents my views. They criticize me for claiming that the extortion problem is merely the monopoly problem in disguise, which I did, and for failing to note that a distinction between these two problems can be made on the basis that extortion uses resources unproductively but monopoly does not, which I did not So that the record on my views may be set straight, I excerpt a few sentences from my earlier paper.

> The problem of "extortion" arises when a change in liability gives rise to a redistribution in wealth In the farmer-rancher case, the relative values of nearby farm and ranchlands will be changed when the rule of liability is altered.. .
>
> In these cases the owner of the specialized resource, ranchland or farmland, that is not required to bear the cost of the interaction may threaten to increase the intensity of the interaction in an attempt to get his neighbor to pay him a larger sum than would ordinarily be required to obtain his cooperation in adjusting the intensity of the interaction downward.... The appropriate economic label for this problem is nothing more nor less than monopoly. It takes on the cast of such legal classifications as extortion only because the context seems to be one where the monopoly return is received by threatening to produce something that is not wanted excessively large herds. The conventional monopoly problem involves a reduction or a threat to reduce the output of a desired good. In the unconventional monopoly problem presented here, there is a threat to increase herd size beyond desirable levels. But this difference is superficial. The conventional monopoly problem can be viewed as one in which the monopolist produces more scarcity than is desired, and the unconventional monopoly problem discussed here can be considered one in which the monopolist threatens to produce too small a reduction in crop damage. Any additional sum that the rancher succeeds in transferring to himself from the farmer is correctly identified as a monopoly return....
>
> Should the law treat such classes of monopoly problem as "extortion"....? .. Activities to which anti-extortion laws normally apply typically involve the use of violence or ... actions considered socially undesirable ... anti-extortion legal measures in such cases [are] less likely to penalize socially desirable actions by mistake. [pp. 22-23]

My identification of the distinction between extortion and monopoly as legal rather than economic stems from an obsolete reluctance to mix normative and positive propositions. The normative judgment as to what is an unproductive use of resources is often more difficult than Daly and Giertz realize. Thus, if A threatens to open a business identical and next door to B's business, it is not generally unlawful for B to pay A to refrain from doing so, whereas if A were to ask payment for not delivering to B's wife a photograph embarrassing to B, then A's activity would generally be held to be illegal. How do the two cases differ? It cannot be said that A's threat, if carried out, has social utility in the first case but not the second. If society frowns upon promiscuous activity, the possibility that

*Professor of economics, University of California-Los Angeles, and senior research fellow, Hoover Institution

such a photograph might be shown to B's wife may deter B from his transgressions, just as, if B should build a railroad next to A's, additional transportation services would be made available to society. Yet in the one case the wealth transfer is legal and in the other it is not.

There are cases in which it is plausible to suppose that the activity clearly yields social cost in excess of social benefit. An example would be the threat to break windows unless paid not to do so. Since the social purposes to be served by window breaking are weak at best, it seems sensible to make wealth transfers brought about through this activity illegal, although some legal government activities, such as payments for not growing crops and for burying pigs to raise pork prices, are highly similar. The more likely it is that the benefits of an activity exceed its cost, the more reluctant we should be to deter, by making illegal, redistributions of wealth associated with the activity.

The general method adopted by the legal system in these situations is essentially to alter the definition of property rights. Thus, firms generally do not own the right to control the kind of business that opens down the street (unless they also own the location down the street). But a person does own the right to some aspects of his privacy, and this is interpreted to mean that he has some control over the photographs taken of him. The legal situation is complex, for the same photograph taken without his permission may be legally taken if it serves the purpose of establishing grounds for divorce. Such a photograph might very well be the evidence used to secure alimony, which, of course, is not extortion because it is legal!

These matters I discussed in a sequel paper (1972b), apparently also not uncovered by Daly and Giertz in their search for my views on extortion.

Even my statements in the comment to which they do refer are inaccurately represented. According to Daly and Giertz, "Demsetz has argued that competition will drive to zero the price of extortionist services and, hence, that it will not be supplied. This confuses cause and effect. It is precisely the fact that extortionist services *are* supplied that drives the price to zero" (p. 1001). What I was discussing was the extortion problem in the context of Coase's assumption of zero transaction costs. According to my article (1971),

> The use of resources to communicate terms of trade or reservation prices or to create an attitude that "he really means it" is unnecessary in a world in which negotiation costs are absent. Shoup's allegation that resource use will be affected by the assignment of liability in a transaction costless world, and, therefore, that the Coase theorem is suspect, is unsubstantiated.... Under competition the price of a good, or of an agreement, is held to its cost. Clearly in this case the production of "excessive" smoke simply to "extort" imposes an unnecessary cost on "extortionists". They would be happy to avoid nonprofitable activity levels for any positive payment... *If negotiating costs are positive, so that resources are used to transfer wealth, there also will be allocative effects.* [pp 444 45, emphasis added]

REFERENCES

G. Daly, and J. F. Giertz, "Externalities, Extortion, and Efficiency," *Amer. Econ. Rev.*, Dec. 1975, *65*, 997-1001.

H. Demsetz, "Theoretical Efficiency in Pollution Control," *Western Econ. J.*, Dec. 1971, *9*, 444-46.

———, (1972a) "When Does the Rule of Liability Matter?," *J. Legal Stud.*, Jan. 1972, *1*, 13.

———, (1972b) "Wealth Distribution and the Ownership of Rights," *J. Legal Stud.*, June 1972, *1*, 223.

Market Efficiency in an Arrow-Debreu Economy: A Closer Look

By KOSE JOHN*

In a recent paper, Mark Rubinstein attempts to develop a precise analytical definition of information efficiency, and to identify the characteristics of participants who believe present security prices fully reflect their information. His model with its surprising implications definitely warrants a closer look.

In Section I, I will present a model of the economy (similar to the one used by Rubinstein but in a somewhat more general context) and derive "nonspeculation conditions" for a general utility function. In Section II, I show that for a special class of utility functions the Rubinstein conditions follow. A summary of his results and their perplexing implications for intertemporal structure of security returns is provided. The role played by the nature of the utility function in arriving at the underlying conditions is highlighted.

I

The model of the economy used is the usual state-preference theoretic framework, with decisions taken over a three-date ($t = 0, 1, 2$) horizon. Let E be an index set for all possible states at $t = 1$, and S at $t = 2$. At date $t = 0$, the individual chooses present consumption C_0 and makes a provisional choice of future consumption by selecting a portfolio of contingent claims $\{\hat{C}_e\}$, ($e \in E$) to $t = 1$ consumption and $\{\hat{C}_s\}$ ($s \in S$) to $t = 2$ consumption. If $\{P_e\}$ and $\{P_s\}$ denote the respective date $t = 0$ prices to these claims, the individual will divide his present (i.e., $t = 0$) wealth W_0 such that

$$W_0 = C_0 + \sum_e P_e\hat{C}_e + \sum_s P_s\hat{C}_s$$

*College of Business Administration, University of Florida. I wish to express my gratitude to Fred D. Arditti and Haim Levy for helpful discussions.

At $t = 1$, some state e is actually realized and contingent claims to that state pay off. The resulting new information revises beliefs held by individuals concerning prices to rule in the market at $t = 2$. In general, the equilibrium prices of contingent claims to date $t = 2$ consumption are revised to $\{P_{s \cdot e}\}$. The individual's wealth W_e at $t = 1$ will therefore depend both on the state e and on his prior choices so that $W_e = \hat{C}_e + \Sigma_s P_{s \cdot e}\hat{C}_s$. In view of the revised prices, each individual with his own revised beliefs will in general desire to revise his provisional choices $\hat{C}_e$ and $\{\hat{C}_s\}$ made at $t = 0$. Let C_e and $\{C_{s \cdot e}\}$ be his revised choices of date $t = 1$ consumption and portfolio of contingent claims to $t = 2$ consumption (which, of course, satisfies his wealth constraint $W_e = C_e + \Sigma_s P_{s \cdot e}C_{s \cdot e}$). At $t = 2$ the true state s occurs, and he consumes $C_{s \cdot e}$.

All individuals obey the Savage axioms of rational choice and aim to maximize expected utility where $\{\pi_e\}$ denotes the probability held by an individual that state e will occur and $\{\pi_{s \cdot e}\}$ denotes the conditional probability that states s will occur given the occurrence of state e. The individual maximizes a general utility function[1] over his final consumption plan $C_0, C_e, C_{s \cdot e}$. The necessary and sufficient first-order conditions for the maximization of expected utility subject to the wealth constraints are used to derive the *nonspeculation conditions*. These are defined as[2] *beliefs and tastes*

[1] It is assumed that the utility function is concave (over the convex feasible set). This ensures that the first-order necessary conditions of optimality are also sufficient.

[2] The definition is patterned after Rubinstein. But in the context of a general utility function, in addition to the beliefs of an individual about the occurrence of states, his tastes for intertemporal consumption (as incorporated in his utility function) would also be involved in his decisions whether or not to revise his portfolio. This modifies results of Rubinstein and other papers which considered the problem for specialized types of utility functions.

for which no portfolio revision is an optimal strategy, that is, given his tastes (as incorporated in the utility function) he perceives the new information that becomes available to him at t = 1 as fully reflected in the revised prices and he opts not to revise his portfolio at t = 1.

The individual's optimization problem is as follows:

$$\max \quad \sum_s \sum_e \pi_e \pi_{s \cdot e} U(C_0, C_e, C_{s \cdot e})$$

with respect to $\quad C_0, \{\hat{C}_e\}\{\hat{C}_s\}\{C_e\}\{C_{s \cdot e}\}$

subject to

$$(1) \quad C_0 + \sum_e P_e \hat{C}_e + \sum_s P_s \hat{C}_s = W_0$$

(2)

$$C_e + \sum_s P_{s \cdot e} C_{s \cdot e} = \hat{C}_e + \sum_s P_{s \cdot e} \hat{C}_s = W_e$$

Forming the Lagrangian function using stochastic Lagrangian multipliers,

$$L = \sum_s \sum_e \pi_e \pi_{s \cdot e} U(C_0, C_e, C_{s \cdot e}) + \lambda_0 \left[W_0 - C_0 - \sum_e P_e \hat{C}_e - \sum_s P_{s \cdot e} \hat{C}_{s \cdot e}\right] + \sum_e \pi_e \lambda_e \left[\hat{C}_e + \sum_s P_{s \cdot e} \hat{C}_s - C_e - \sum_s P_{s \cdot e} C_{s \cdot e}\right]$$

From here on, I will denote the optimal values $C_0^*, C_e^*, C_{s \cdot e}^*$ of the arguments C_0, $C_e, C_{s \cdot e}$ of the U-function and its partials by three dots. Further denote

$$\partial U/\partial C_0(\ldots) = U_0(\ldots), \partial U/\partial C_e(\ldots) = U_1(\ldots), \partial U/\partial C_{s \cdot e}(\ldots) = U_2(\ldots)$$

Writing the first-order necessary conditions of optimality,

$$(3) \quad \frac{\partial L}{\partial C_0} = \sum_e \sum_s \pi_e \pi_{s \cdot e} U_0(\ldots) - \lambda_0 = 0$$

(4)

$$\frac{\partial L}{\partial C_e} = \pi_e \sum_s \pi_{s \cdot e} U_1(\ldots) - \lambda_e \pi_e = 0 [\forall e \in E]$$

$$(5) \quad \frac{\partial L}{\partial C_{s \cdot e}} = \pi_e \pi_{s \cdot e} U_2(\ldots) - \lambda_e \pi_e P_{se} = 0 \; [\forall e \in E, \forall s \in S]$$

$$(6) \quad \frac{\partial L}{\partial \hat{C}_e} = -\lambda_0 P_e + \pi_e \lambda_e = 0 \; [\forall e \in E]$$

(7)

$$\frac{\partial L}{\partial \hat{C}_s} = -\lambda_0 P_s + \sum_e \pi_e \lambda_e P_{se} = 0 [\forall s \in S]$$

and the feasibility conditions (1) and (2). Rewriting (3) and (4)

$$(8) \quad \lambda_0 = \sum_e \sum_s \pi_e \pi_{s \cdot e} U_0(\ldots)$$

$$(9) \quad \lambda_e = \sum_s \pi_{s \cdot e} U_1(\ldots) \; [\forall e \in E]$$

Rewriting (5) (6) and (7) we have

$$(10) \quad \lambda_e = \frac{\pi_{s \cdot e}}{P_{s \cdot e}} U_2(\ldots)$$

$$(11) \quad \frac{\lambda_0}{\lambda_e} = \frac{\pi_e}{P_e}$$

$$(12) \quad \lambda_0 P_s = \sum_e \lambda_e \pi_e P_{s \cdot e}$$

Multiplying (10) by (11) we obtain

(13)

$$\frac{\lambda_0}{U_2(\ldots)} = \frac{\pi_e \pi_{s \cdot e}}{P_e P_{s \cdot e}} \quad [\forall s \in S, \forall e \in E]$$

After the state e is realized and the subsequent revision in prices and beliefs, if $C_e^* = \hat{C}_e^*$ and $\{C_{s \cdot e}^*\} = \{\hat{C}_s^*\}$ satisfy (13) for all states s, then "no portfolio revision" would be an optimal strategy.[3] In other

[3]We can easily verify that such choices are feasible at date t = 0. Substituting $\lambda_e = \lambda_0(P_e/\pi_e)$ from (11) in (12) we obtain

$$(a) \quad P_s = \sum_e P_e P_{s \cdot e}$$

Multiplying (2) by P_e and summing over e,

$$\sum_s \left(\sum_e P_e P_{s \cdot e}\right) C_{s \cdot e}^* + \sum_e P_e C_e^* = \sum_s \left(\sum_e P_e P_{s \cdot e}\right) \hat{C}_s^* + \sum_e P_e \hat{C}_e^* \text{ and}$$

words if the initial choices were such that (13) holds for the revised beliefs and prices, then the currently held portfolio is optimal.

The main implication of this condition (13) is that tastes of an individual for intertemporal consumption (in addition to beliefs and prices) is involved in his decision to revise his portfolio. Two individuals having the same beliefs about the incidence of states (and of course confronted with the same set of prices) could take different decisions about revising their portfolios.

Another immediate implication is that the utility-free nonspeculation conditions (derived by Rubinstein) would not hold[4] for a general utility function. But interestingly such conditions do hold for a class of utility functions (discussed in Section II) which extends the Rubinstein class of additively separable functions.

II

In this section I examine a class of utility functions for which a utility-free nonspeculation condition holds and define "E-S separable functions" as a class of utility functions $U(C_0, C_e, C_{s\cdot e})$ which could be separated as $V(C_0, C_e) + W(C_0, C_{s\cdot e})$. This class of utility functions separate (additively) the utilities of consumption in period t = 1 and t = 2; the marginal utility of t = 2 consumption is independent of the consumption in (t = 1) state e.

THEOREM: *If the utility function is E-S separable then the nonspeculation condition is*

$$\frac{\pi_e \pi_{s\cdot e}}{P_e P_{s\cdot e}} = \frac{\pi_s}{P_s} \tag{14}$$

PROOF:

Since U is E-S separable,

$$U(C_0, C_e, C_{s\cdot e}) = V(C_0, C_e) + W(C_0, C_{s\cdot e})$$

$$U_2(\ldots) = \frac{\partial W}{\partial C_{se}}(C_0^*, C_{s\cdot e}^*) = W_2(C_0^*, C_{s\cdot e}^*) \tag{15}$$

Substituting (15) in (13) we have

$$\frac{\pi_e \pi_{s\cdot e}}{P_e P_{s\cdot e}} = \frac{\lambda_0}{W_2(C_0^*, C_{s\cdot e}^*)} \tag{16}$$

If $C_{s\cdot e}^*$ is independent of e, the right-hand side of (16) is independent of e. Then there exists a number q_s (independent of e) such that

$$q_s = \frac{\pi_e \pi_{s\cdot e}}{P_e P_{s\cdot e}} = \frac{\lambda_0}{W_2(C_0^*, C_{s\cdot e}^*)}$$

Summing over the states e,

$$q_s\left(\sum_e P_e P_{s\cdot e}\right) = \sum_e \pi_e \pi_{s\cdot e}$$

which by equation (a) (in fn. 3) implies $q_s P_s = \pi_s$, i.e.,

$$q_s = \frac{\pi_s}{P_s} = \frac{\pi_e \pi_{s\cdot e}}{P_e P_{s\cdot e}} = \frac{\lambda_0}{W_2(C_0^*, C_{s\cdot e}^*)} \tag{17}$$

Conversely if (17) holds

$$\frac{\lambda_0}{W_2(C_0^*, C_{s\cdot e}^*)} = \frac{\pi_s}{P_s}$$

so that $C_{s\cdot e}^*$ must be independent of e. Therefore we can define C_s^* so that $C_{s\cdot e}^* = C_s^*$ if and only if (14) holds. To show $\{C_s^*\} = \{\hat{C}_s^*\}$ and $\{C_e^*\} = \{\hat{C}_e^*\}$ we just need to verify that these choices are feasible at date t = 0. (This is done as in fn. 3.) In this case the individual would be willing to move immediately to $\{C_s^*\}$ and $\{C_e^*\}$ at date t = 0, and not revise his portfolio at t = 1.

COROLLARY: *For an additively separable utility function,*

$$U(C_0, C_e, C_{s\cdot e}) = P(C_0) + Q(C_e) + R(C_{s\cdot e})$$

Using (a) and (1)

$$\sum_s P_s C_{s\cdot e}^* + \sum_e P_e C_e^* = \sum_s P_s \hat{C}_s^* + \sum_e P_e \hat{C}_e^* = W_0 - C_0^*$$

i.e., $\Sigma_s P_s C_{s\cdot e}^* + \Sigma_e P_e C_e^* + C_0 = W_0$, the feasibility condition at t = 0.

[4]Consider equation (13). Since $U_2(\ .\ )$ contains as its argument C_s^*, the left-hand side of (13) depends on e, even if $C_{s\cdot e}^*$ is stipulated to be independent of e. The condition that $\lambda_0/U_2(.\ .)$ be independent of e is critical in the Rubinstein proof for his nonspeculation condition (see Rubinstein, p. 816).

the nonspeculation conditions are the same, i.e., (14).

PROOF:

Additively separable functions form a proper subclass of the *E-S* separable functions.

For this class of functions the nonspeculation conditions depend only on the relationships between beliefs and prices. Rewriting (14) we have

$$\frac{P_{s,e}/P_s}{\pi_{s,e}/\pi_s} = \frac{\pi_e}{P_e} \tag{18}$$

The right side is fully determined before e is known. It is a constant of proportionality between revised prices and beliefs predetermined at $t = 0$ which must hold for "no revision" to be optimal.

Now we look at some restrictions on the intertemporal structure of security returns implied by the nonspeculation condition (14). Translating prices into rates of return,

$$1 + r_{1s,e} = P_{s,e}/P_s \tag{19}$$

$$1 + r_{1f} = (\Sigma_e P_e)^{-1} \tag{20}$$

$$1 + r_{2s,e} = P_{s,e}^{-1}, \text{etc} \tag{21}$$

the nonspeculative intertemporal structure follows from the nonspeculation condition, as

$$E[(1 + r_{1s})(1 + r_{2s})] = (1 + r_{1f})E[(1 + r_{2s})] \tag{22}$$

where the expectations are assessed with respect to beliefs held at $t = 0$. In words, the intertemporal structure of rates of return implied by nonspeculative beliefs has the following characterization:[5] the expected one plus compound rate of return on any security discounted by one plus the first-period risk-free rate equals its expected one plus second-period rate of return where the expectations are assessed with respect to beliefs at the beginning of the first period. If we accept this characterization of the intertemporal structure some popular statistical hypotheses used in this context lead to perplexing conclusions. Serially uncorrelated rates of return, (as, for example, surveyed in Eugene Fama) have played a significant role in the guise of the random walk hypothesis in statistical tests of information efficiency. But with serially uncorrelated rates,

$$E[(1 + r_{1s})(1 + r_{2s})] = E[(1 + r_{1s})]E[(1 + r_{2s})] \tag{23}$$

is clearly equivalent to

$$E[(1 + r_{1s})] = 1 + r_{1f} \tag{24}$$

using (22). That is, the expected first-period rate of return on any security equals the first-period risk-free rate if and only if its intertemporal sequence of one plus rates of return are serially uncorrelated. This result is counterintuitive in a risky economy composed of risk-averse individuals.[6]

In conclusion, even though the model looks promising, its implications based on a special class of utility functions warrant closer scrutiny. Two immediate directions for further research suggest themselves: 1) to identify alternate conditions for the general utility function and to examine their implications; and 2) to recognize classes of utility functions and the specific nonspeculation conditions they would imply.

[5]This characterization will in general hold only for rates of return of each contingent claim, not for rates of return on portfolios of contingent claims.

[6]The role of the nature of the utility function (i.e., its *E-S* separability) in arriving at the underlying condition (14) has been discussed. So if the participants of an economy cannot be characterized by *E-S* separable utility functions then (14) and its implications need not hold even if they perceive the securities market as informationally efficient.

REFERENCES

K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev Econ. Stud.*, Apr. 1964, *31*, 91–96.

E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, *25*, 383–417.

M. Rubinstein, "Securities Market Efficiency in an Arrow-Debreu Economy," *Amer. Econ. Rev.*, Dec. 1975, *65* 812–24.

P. A. Samuelson, "Proof that Properly Anticipated Prices Fluctuate Randomly," *Ind. Manage. Rev.*, Spring 1965, *6*, 41–50.

Money, Income, and Causality in the United States and the United Kingdom: A Theoretical Explanation of Different Findings

By BLUFORD H. PUTNAM AND D. SYKES WILFORD*

In an article in this *Review*, Christopher Sims presented an innovative statistical technique to determine the direction of causality, then applied this methodology to money and nominal income in the United States. He concluded: "The main empirical finding is that the hypothesis that causality is unidirectional from money to income agrees with the postwar *U.S.* data, whereas the hypothesis that causality is unidirectional from income to money is rejected" (p. 540). In a more recent paper in this *Review*, David Williams, C. A. E. Goodhart, and D. H. Gowland applied Sims' statistical methodology to the United Kingdom and concluded: "We found for the *U.K.* some evidence of unidirectional causality running from nominal incomes to money but also some evidence of unidirectional causality running from money to prices. Taken together, this evidence suggests, perhaps, a more complicated causal relationship between money and incomes in which both are determined simultaneously" (p. 423). Furthermore, Williams, Goodhart, and Gowland suggest some general possibilities for the differences between the United States and the United Kingdom, and they are careful to note that: "Because of the various differences in context the finding that in the United Kingdom the relationship between money and income appears different from that found by Sims for the United States in no way casts any doubt on the validity of Sims' own results" (p. 417).

The purpose of this paper is to present a concise model which draws together the findings of Sims for the United States and Williams, Goodhart, and Gowland for the United Kingdom. To accomplish this, a fixed exchange rate system is modeled in which one country, the United States, serves as the primary reserve currency country, while other countries, the United Kingdom in this case, hold a substantial portion of their international reserves denominated in terms of the reserve currency.[1] Particular attention is paid to the asymmetrical nature of the system with respect to money's influence on nominal income and vice versa. Indeed, the ability of the reserve currency country to create international reserve assets plays the primary role in explaining the asymmetrical nature of the system and the empirical results of Sims, and Williams, Goodhart, and Gowland.

The model is couched in a world in which asset reallocations are viewed as adjustments toward maintaining general equilibrium. This equilibrium is based on a stable set of preferences regarding the structure of individual portfolios, broadly defined in terms of holdings of real consumption goods, real interest bearing financial assets, real money balances, and leisure time. The assumption of equilibrium conditions in all markets allows attention to focus directly on the money market to isolate the process of portfolio adjustment in international markets. As in similar models based on the monetary approach to

*Economists, Chase Manhattan Bank, N.A. The views expressed in this paper are solely our own and do not necessarily represent those of the Chase Manhattan Bank. We wish to thank David T. King, the managing editor of this *Review*, J. Richard Zecher, J. E. Tanner, Walton T. Wilford, C. A. E. Goodhart, and Marc A. Miles for their comments on earlier drafts.

[1]One may note that for the Commonwealth countries the British pound acted as a reserve currency. However, the pound's relative world influence vis-à-vis the *U.S.* dollar was small during the Bretton Woods period.

the balance of payments, three equations are specified for each country, a money demand function, a money supply identity, and the condition of monetary equilibrium.[2] The difference between the reserve currency country and all other countries lies in the money supply identity. To follow through the consequences of the differences, the case of the reserve currency country is presented first, and then the case for other countries.

I. The Reserve Currency Country

The model of the reserve currency country contains three straightforward equations. For simplicity, money demand depends on permanent real income and the price level:

$$(1) \qquad M_A^d = k_A P_A Y_A$$

where the subscript A stands for the reserve currency country, and

k = constant
P = price level
Y = permanent real income
M^d = quantity of money demanded

The reserve currency country's central bank has the power to determine the nominal money stock. Thus

$$(2) \qquad M^s = M$$

This money supply equation does not ignore the fact that the reserve currency country operates in the context of international markets (i.e., is an open economy). Equation (2) implies that the domestic money stock of the reserve currency country need not be responsive to its balance of payments. When money flows from the reserve currency country to the rest of the world, that money eventually is purchased by foreign central banks, due to their agreement to buy and sell currency at a fixed price. However, foreign central banks do not hold the major portion of their international reserves in noninterest-bearing form when an alternative exists. In this case, the foreign central banks purchase government securities from the reserve currency country.[3] Under these conditions, a balance-of-payments deficit results in an outflow of government securities, but not of money. Thus, the central bank of the reserve currency country maintains control over the nominal money stock. Furthermore, the reserve currency country is the only country with the power to create international reserves, thereby enabling it to remain independent of external influences.

Finally, the model is closed with the condition of monetary equilibrium:

$$(3) \qquad M^d = M^s$$

Solving this system for the price level and converting to growth terms yields

$$(4) \qquad g(P_A) = g(M_A) - g(Y_A)$$

where $g(X) = (dX/dt)/X$, for $X = P, M$, and Y.

Growth of the price level is determined by the growth of the money supply relative to permanent real income growth. If permanent real income is growing at a stable rate, and is for the most part independent of monetary disturbances, then changes in nominal income are caused primarily by changes in the supply of money the result which Sims obtained for the United States. The essential point is that it is not necessary to postulate that the United States is a

[2]There are several articles which summarize the essentials of the monetary approach to the balance of payments. For particular references see Harry Johnson (1973, 1976); Jacob Frenkel and Johnson, Frenkel and Carlos Rodriguez, Donald Kemp, Putnam, Wilford, J. Richard Zecher

[3]During the Bretton Woods period, when a particular central bank received dollars, which it held as international reserves, a portion of these dollars, that which was not needed for governmental transactions demand, were used to buy U.S. government securities. That is, the dollars returned to the United States, and instead of holding currency, foreign central banks held U.S. government securities. Thus, one may equate the flow of a dollar from a foreign central bank to the United States, not to a loss of international reserves for the United States, but to a securities account transaction between the authorities of the United States and the foreign central bank. As such, this transaction would be counted in the U.S. official settlements balance, since U.S. liabilities to foreign official agencies are affected. But, the money supply of the United States is not necessarily affected. For a more detailed explanation of this point see Kemp (pp. 17–20), Alexander Swoboda (p. 17), or Lance Girton and Dale Henderson.

closed economy to explain Sims' results. During the Bretton Woods period, the dollar played a leading role in world economic activity. However, by serving as the reserve currency country, the United States maintained control over its domestic money supply.

II. The Rest of the World

The asymmetrical properties of the international monetary system become apparent when examining the nonreserve currency case. For simplicity, the money demand function is unchanged from the preceding analysis:

$$M_j^d = k_j P_j Y_j \tag{5}$$

where the subscript j stands for any nonreserve currency country.

With respect to the money supply identity, however, there are major differences between the reserve currency country and other countries. In the process of maintaining fixed exchange rates, central banks will incur inflows and outflows of reserve currency assets (i.e., international reserves) Because their monetary base depends on the asset portfolio of the central bank, the balance of payments will directly affect the domestic money supply. That is:

$$M_j^s = aH = a(R + D) \tag{6}$$

where a = the money multiplier
H = high-powered money, or the monetary base
R = international reserve assets (liabilities of the reserve currency country)
D = domestic assets held by the central bank

Dividing high-powered money, which makes up the bulk of the liability side of the central bank's portfolio, into international reserves and domestic assets from the asset side of the portfolio, introduces the effects of the balance of payments directly into the money supply identity.[4] This formulation of high-powered money allows one to focus directly upon the sources of high-powered money, foreign or domestic.

The condition of monetary equilibrium completes the model:

$$M_j^s = M_j^d \tag{7}$$

Solving the system for the international reserve flow and converting to growth terms yields

$$\frac{R}{H} g(R) = g(P_j) + g(Y_j) - g(a) - \frac{D}{H} g(D) \tag{8}$$

In this monetary approach to the balance of payments, the nonreserve currency country's prices are determined on unified world markets and are given exogenously to the domestic economy. The balance of payments reflects attempts by individuals to maintain equilibrium money balances as they adjust their expenditures and receipts. The endogeneity of the international reserve flow thereby implies the endogeneity of the money supply, leading to the conclusion that the nonreserve currency country's monetary authorities cannot determine the domestic money stock.

Two asymmetrical properties of the international system are now apparent. First, the reserve currency country can control its money supply, while other countries cannot. Secondly, the reserve currency country can influence its price level, while other countries must accept prices as determined on unified world markets. This holds true even though the reserve currency country participates in the same world markets. To illustrate, suppose from a position of equilibrium, the reserve currency country expands its money supply. Prices begin rising in the reserve currency country and also simultaneously in world markets. Other countries trading in these markets experience a rising price level. Nominal income increases, and there is an inflow of international reserves as individuals seek to maintain equilibrium money balances. The balance-of-payments surplus reestablishes monetary equilibrium by expanding the

[4]A detailed explanation of this view of the money supply process can be found in Zecher, Kemp, or Leonall Andersen and Jerry Jordan.

supply of money to meet demand. Now suppose that from a position of equilibrium a nonreserve currency country attempts to raise its nominal money supply. As individuals reduce their excess money balances, the central bank experiences an outflow of international reserves due to residents' increased expenditures on foreign goods and assets. The loss of international reserves impacts directly on the monetary base, and the money supply moves back toward its initial level. International reserve flows equilibrate money supply to money demand as determined primarily by permanent real income and the price level. With prices determined on unified world markets and permanent real income not directly affected by the balance of payments, then essentially, nominal income and the money supply are simultaneously determined—a conclusion supported by the empirical findings of Williams, Goodhart, and Gowland for the United Kingdom.[5]

III. Summary

A fixed exchange rate system in which one country serves as the reserve currency country has important asymmetrical properties. Indeed, only the reserve currency country can control its money supply. From this property, several implications concerning the direction of causality follow directly. Control of the money supply results in the ability to influence the price level, and thus nominal income in the reserve currency country. Hence, Sims found that causality flows from money to nominal income in the United States. Furthermore, changes in prices and nominal incomes in the reserve currency country will simultaneously affect conditions in world markets. Individuals in other countries, reacting to these changes, adjust their portfolios. This adjustment process prompts simultaneous changes in prices, nominal income, and the money stock in nonreserve currency countries. Thus, Williams, Goodhart, and Gowland found that in the United Kingdom neither money or nominal income cause each other.

Both Sims' and Williams, Goodhart, and Gowland's investigations covered the Bretton Woods period, during which the world was essentially operating on a dollar standard. Given a model highlighting the asymmetrical properties of such an international monetary system, the apparently inconsistent findings by Sims for the United States and Williams, Goodhart, and Gowland's for the United Kingdom are instead entirely compatible.

[5]The reader should note that the empirical technique employed by both Sims, and Williams, Goodhart, and Gowland can only indicate causality when leads and lags exist among the variables studied. For clarity, no attempt has been made to specify these leads and lags mathematically, but as the preceding paragraph indicates, there is a clear presumption as to their existence and direction.

REFERENCES

L. Andersen and J. Jordan, "The Monetary Base: Explanation and Analytical Use," *Fed. Reserve Bank St. Louis Rev.*, Aug. 1968, *50*, 7-11.

J. Frenkel and C. A. Rodriguez, "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674-88.

——— **and H. G. Johnson,** "The Monetary Approach to the Balance of Payments: Essential Concepts and Origins," in their *The Monetary Approach to the Balance of Payments*, Toronto 1976.

L. Girton and D. W. Henderson, "Financial Capital Movements and Central Bank Behavior in a Two Country, Short-Run Portfolio Balance Model," *J. Monet. Econ.*, Jan. 1976, *2*, 33-61.

Harry G. Johnson, *Further Essays in Monetary Economics*, Cambridge 1973, 229-49.

———, "Elasticity, Absorption, Keynesian Multiplier, Keynesian Policy, and Monetary Approaches to Devaluation Theory: A Simple Geometric Exposition," *Amer. Econ. Rev.*, June 1976, *66*, 448-52.

D. S. Kemp, "A Monetary View of the Balance of Payments," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1975, *57*, 14-22.

B. H. Putnam, "Non-traded Goods and the Monetary Approach to the Balance of

Payments," res. paper no. 7714, Fed. Reserve Bank, New York 1976.

C. A. Sims, "Money, Income, and Causality," *Amer. Econ. Rev.*, Sept. 1972, *62*, 540-52.

A. K. Swoboda, "Gold, Dollars, Eurodollars, and the World Money Stock under Fixed Exchange Rates," *Amer. Econ. Rev.*, Sept. 1978, forthcoming.

D. Sykes Wilford, *Monetary Policy and the Open Economy: Mexico's Experience*, New York 1977.

D. Williams, C. A. E. Goodhart, and D. H. Gowland, "Money, Income, and Causality: The U.K. Experience," *Amer. Econ. Rev.*, June 1976, *66*, 417 13.

J. R. Zecher, "Monetary Equilibrium and International Reserve Flows in Australia," *J. Finan.*, Dec. 1974, *29*, 1523-530.

Currency Substition, Flexible Exchange Rates, and Monetary Independence

By MARC A. MILES*

In the persistent fixed vs. flexible exchange rate debate, one of the most common arguments in favor of flexible exchange rates is that they insulate a country's money supply from monetary developments in the rest of the world (see, for example, Milton Friedman, 1953; Robert Mundell). Under fixed rates such monetary independence is impossible because, by pegging the value of domestic currency to foreign currency, the central bank makes foreign currency a perfect substitute for domestic currency on the supply side. Should the monetary authorities in country *A* increase the money supply once and for all, the domestic money supply would exceed domestic money demand, and money would immediately flow out through the balance of payments. The domestic balance-of-payments deficit must be matched by a balance-of-payments surplus abroad. Thus money supplies abroad must also increase, and a common rate of inflation would be observed among countries. But perfectly flexible exchange rates are assumed to eliminate this source of monetary interdependence. Under flexible rates the balance of payments is always zero, that is, there is no net money flow between central banks. Central banks are no longer allowed to intervene to guarantee the value of their currencies Thus flexible rates make currencies perfect nonsubstitutes on the supply side.

The lack of intervention by the central banks and the accompanying elimination of the substitution of currencies on the supply side in turn is assumed to result in no net movements of money among countries at all and thus complete monetary independence. But this monetary independence argument makes the implicit assumption that currencies are also nonsubstitutes on the demand side, that is, Frenchmen hold only francs and Germans only deutschemarks. It is assumed that no foreign currency is held by domestic transactors for either transactions, speculative or precautionary purposes. However, in the context of the existing international economic environment this assumption appears quite dubious. Multinational corporations have strong incentives to diversify the currency composition of their cash balances in order to facilitate their endeavors in various countries. Even individuals and businesses that are clearly domiciled in a particular country often have transactions or precautionary or even speculative motives for diversifying the currency composition of their money holdings. Anyone who consistently makes purchases from foreign countries has at least the same transactions motives for demanding foreign currency balances as for demanding domestic currency balances. Importers and exporters, businessmen who travel abroad, tourists, and residents of border areas all have incentives to diversify their currency balances By holding foreign money, the transactions costs of their foreign purchases are reduced. With a significant subset of a country's citizens and businesses maintaining diversified currency portfolios, the conclusion of independent monetary policy no longer appears valid. A change in a country's monetary policy can generate an adjustment in the currency composition of cash balances and thus an intercountry movement of currencies which can offset at least part of the policy change.

In this paper the question of currency substitution is examined. In Section I the possible mechanisms through which the

*Rutgers College. I would like to thank John Van Belle, Arthur Laffer, and William Gasser for their help in developing this paper. Useful comments were also provided by an anonymous referee. Sarah Biser provided research assistance.

substitution can occur are discussed. Two mechanisms are presented, corresponding to whether an increase in the money supply causes a drop in the interest rate or a rise in inflationary expectations. Section II shows the implications of currency substitution on the existence of independent monetary policies under perfectly flexible exchange rates. The conclusion is that where currency substitution exists, even perfectly flexible rates may not guarantee monetary independence. In Section III a constant elasticity of substitution (*CES*) production function for the services of money is used to derive a testable model of currency substitution. This model is then tested on Canadian data. It is found that a high degree of currency substitution exists in Canada, especially during floating rate periods. The paper concludes with a summary of the implications.

I. The Mechanism of Currency Substitution

The mere ownership of foreign currency-denominated balances by domestic residents is not a sufficient condition for currency substitution to occur. A given amount of foreign currency balances may exist within the country for institutional or historical reasons. For currency substitution to exist, not only must there be foreign currency balances, but the level of these balances must change in response to changes in other economic variables. Furthermore, it is not a necessary condition for currency substitution that each individual within the country hold foreign currency balances. Currency substitution requires only that there exist a group of individuals who, given the current value of economic variables, hold both domestic and foreign currency balances and are indifferent at the margin between holding more domestic or more foreign balances.

In order to discuss the mechanism of currency substitution, assume that such a group of individuals exist. These individuals may be foreign traders, border residents, or even multinational corporations. The important characteristic of each, however, is that they hold a diversified portfolio of real money balances. While the overall size of the real cash balance portfolio will vary with the level of real income and the returns on other types of assets, the composition of the portfolio will vary with the relevant opportunity costs of holding real balances of the various types of currencies. If the opportunity cost of holding real balances denominated in currency A rises relative to the opportunity cost of holding those denominated in currency B, all of these individuals will be assumed to reduce their real balances denominated in currency A and to increase their holdings denominated in currency B.

The particular aspect of currency substitution that is of interest here is its effect on monetary policy. The discussion of the mechanism of currency substitution will therefore be in terms of the interaction of monetary policy and currency substitution. Specifically we will examine whether changes in monetary policy change relative costs of holding currency and thus induce offsetting inflows or outflows of money.

The model will be assumed to consist of two countries A and B, each supplying its own domestic currency-denominated money asset, C_A and C_B, respectively. Both countries are assumed to have a group of individuals who hold real balances denominated in both C_A and C_B in their cash balance portfolios. The ratio in which the jth individual holds real balances of the two currencies can be described by

$$m_j(r) = (c_A/c_B)_j \tag{1}$$

where $c_A = C_A/P_A$ is the level of real balances denominated in currency A, and $r = i_B/i_A$ is the ratio of the opportunity costs of holding real balances in C_B and C_A. As r rises, m_j is also assumed to rise. The model is assumed initially at equilibrium, so that each individual's total demand for real balances precisely equals his holdings. Also, given the prevailing opportunity costs of holding c_A and c_B, each individual is assumed to have adjusted his cash balance portfolio so that he is just indifferent between holding a little more c_A or a little more c_B.

Now suppose that the monetary authorities of A increase the supply of C_A, the supply of C_B remaining constant. Following Friedman (1969) there are two possibilities for this increase, a once and for all increase or a change in the rate of increase from zero to a positive number. The two possibilities will be considered in turn. In both cases it will initially be assumed that the entire adjustment to the change in the quantity of C_A occurs through a change in P_A. This assumption will be subsequently relaxed.

A. *A Once and for All Increase in* C_A

Assume, as does Friedman, that the monetary authorites of country A unexpectedly decide to perform a once and for all increase in the quantity of C_A by dumping from a helicopter additional units of C_A. For ease of exposition it will be assumed that each individual's cash balances increase in proportion to his initial holdings. If the economy was initially at equilibrium, this monetary shock will require adjustments in order to return the economy to equilibrium. In terms of the present model possible adjustments of four variables are of interest. First, how does the monetary shock affect the rate of interest i_A? Second, how does the change in i_A affect desired ratios of real balances m? Third, what change in P_A is required to achieve equilibrium? Finally, do changes in the preceding three variables cause a redistribution of C_A between the two countries? Each of these questions will be answered in turn.

If there is a once and for all increase in C_A, the real quantity of money assets denominated in C_A increases relative to the real value of other assets denominated in C_A. In order for the additional relative amounts of c_A to be absorbed, the cost of borrowing c_A balances must fall. Thus the increase in C_A causes i_A to fall. However, since it is assumed that the real value of C_B and all other assets denominated in C_B remain constant, the cost of borrowing c_B balances remains unchanged. The value of r is observed to rise.

The rise in r makes holding relatively more c_A suddenly more attractive. Thus m will also rise. But m does not rise only in country A. Since individuals anywhere in the world are assumed to face the same opportunity cost of holding real balances denominated in a particular currency, r rises in both countries. Thus the new equilibrium must be characterized by a higher value of m in both countries as compared to the initial equilibrium.

Two other conditions must also hold in the new stock equilibrium. First, in each country the supply of C_A must be equal to the demand for C_A. This condition will be satisfied when supply of c_A equals the demand for c_A in each country. Second, if money demand equals money supply in each country, then total world demand must equal total world supply. These two conditions will determine the final rise in P_A and the distribution of C_A among countries.

For example, for total world money demand to equal money supply, the excess world supply of real balances denominated in C_A must be eliminated. One possibility is for P_A to rise sufficiently to completely eliminate the excess supply in country A. This is the solution suggested by flexible exchange rate models. The derivation of this solution is the fact that all the increase in the supply of C_A occurs in country A. But while all the increase in supply occurs in A, all the increase in demand does not. The desired value of m rises in both countries, implying that at even a constant P_A there is excess demand for C_A in country B. Thus while such a rise in P_A will eliminate the excess supply in A, it will only increase the excess demand in B. Since such a price rise does not completely eliminate the world excess demand for C_A, it cannot be the equilibrium price rise. Obviously world equilibrium requires a smaller rise in P_A. But with a smaller price rise, there will be excess supply of C_A in country A. The final equilibrium rise in P_A will therefore be determined by where the excess supply of C_A in A is precisely equal to the excess demand for C_A in B.

While the rise in P_A is sufficient to deter-

mine world equilibrium in the C_A market, it has not equated the markets within the two countries. The inability for changes in P_A to equate both world and domestic markets is caused by the distribution effect that demand has risen in both countries while supply has risen in only one. The only way for domestic markets to clear, given the P_A that creates world equilibrium, is for units of C_A to flow from country A, where there is excess supply, to country B where there is excess demand. Thus a once and for all increase in C_A in country A has caused units of C_A to flow between countries.

B. *An Increase in the Rate of C_A Increase from Zero to a Positive Number*

Now assume that the monetary authorities of country A decide to send the helicopter over the country at regular intervals. The monetary authorities inform the public of this decision and that each trip of the helicopter will increase the amount of C_A by a constant percentage. Again, the changes in r, m, P_A, and the distribution of C_A between countries will be of interest.

In contrast to the once and for all increase, the decision to increase the supply of C_A at a constant rate will create expectations of inflation which will cause i_A to rise. The rise in i_A will cause r to fall. The fall in r will cause the desired m to fall in both countries as individuals try to reduce the share of real balances denominated in C_A.

The ensuing rise in P_A in each period can be divided into two parts, the rise in P_A necessary to maintain the initial level of relative real balances in each country and the rise in P_A that reduces m from its initial level. The second source of a rise in P_A will have a positive value in the first period, and a zero value in subsequent periods. For example, consider the first period. The supply of C_A has been increased in country A. Ignoring for a moment any change in desired m, the analysis is very similar to the once and for all increase case. The level of P_A must rise in order to equilibriate the world supply and demand for C_A. But in order for the supply of C_A to equal demand in both countries, units of C_A must flow from A to B. Again the increase in C_A causes a flow of currency between countries.

But the expectation of inflation causes i_A to rise, and m will not remain constant. The value of m must fall in both countries. Assuming symmetrical responses, the desired level of m will fall by the same percentage in the two countries. This adjustment can be accomplished by a once and for all rise in P_A equal to the percentage fall in m. Such a rise will reduce c_A in both countries by the desired amount and will not require further flows of C_A between the countries.

So only one of the two rises in P_A causes net flow of C_A in the first period, and the flow is again from country A to country B.[1] Furthermore, once the level of m adjusts to the level consistent with the expected inflation, there will no longer be any force creating a second source of a rise in P_A. In subsequent periods P_A will adjust only to maintain real balances at the initial m level, and as has been shown, this adjustment requires a movement of C_A from A to B. So a continuous increase in C_A produces offsetting outflows of C_A in each period.

[1]The analysis has concentrated exclusively on the substitution effect of inflationary expectations. However, there is another effect, the change in the total demand for real balances relative to the total portfolio supply. Inflationary expectations on the one hand reduce the total demand for real balances by raising the weighted average of the cost of holding real balances, a point emphasized in the description of the empirical model. On the other hand, inflation reduces the real supply of C_A in any country holding it. Naturally those countries holding c_A in the greatest proportion will find their supply of real balances reduced by the largest percentage. It is at least conceivable that this change in total real balance demand relative to supply could cause the opposite type of money flow to occur. For example, with hyperinflation in country A, the large rise in P_A will reduce the total value of real balances in country A more than any other country. If that supply is falling faster than demand, country A will have to be a net importer of money, and there will be a net flow of money from country B to country A, the opposite of the flow described above. In this particular case the overall demand for balances will dominate the substitution effect. However, while this case is possible, in most cases the substitution effect should dominate.

C. Allowing P_B to also Adjust

Until now it has been assumed that when C_A increases all the adjustment in the level of real balances occurs through a rise in P_A. However, if currencies are substitutable, it is quite possible that at least some of the adjustment can occur through a rise in P_B. When C_A is increased by the monetary authorities in A, the overall level of real balances in the world is increased above the level of world demand. To this point it has been assumed that the overall level is reduced back to an equilibrium level by lowering only c_A. But lowering c_B through a rise in P_B also has this desired effect. In addition lowering c_B can also have the desired effect on m. For example, in the case of the once and for all rise in C_A, a rise in P_B will help to cause m to rise to the desired level and P_A will not have to rise as much as in the previous case.

II. Implications for Flexible Exchange Rates and Monetary Independence

Once the assumption of currency substitution on the demand side is introduced, the conclusion that perfectly flexible exchange rates imply independent monetary policy begins to evaporate As shown above, currency substitution in demand produces flows of money and changes in price levels that are not consistent with the traditional flexible exchange rate model When the monetary authorities of A increase the supply of C_A, rather than the entire increase remaining within country A and P_A adjusting to eliminate the domestic excess supply of real balances, some units of C_A are redistributed through private markets to country B. The effects of the monetary policy are no longer internalized within A. Without any intervention by governments of either country A or country B, the nominal money supply also rises in B.

Once possible changes in P_B are introduced, the implications become even more obvious. As long as c_A and c_B are substitutes in demand, a rise in C_A can cause not only P_A to rise, but also P_B. Thus inflation is transmitted between countries without having to assume any government intervention in the foreign exchange market. Yet transmission of inflation is precisely the type of phenomenon from which flexible rates are assumed to insulate a country.

The degree to which inflation will be transmitted between the countries will of course be proportional to the degree of substitution between currencies. The limiting case is where c_A and c_B are perfect substitutes. In that case there is the equivalent of one world currency, just as when central banks make currencies perfect substitutes on the supply side by fixing exchange rates. In that case no distinction can be drawn between either c_A or c_B, or between P_A or P_B. An increase in the nominal money supply in either country will increase the real balances used in both countries and cause the price level in both countries to rise by precisely the same amount. The degree of substitution in demand between currencies therefore becomes an important empirical question.

III. An Empirical Model of Currency Substitution

For the most part the concepts of currency substitution and the diversification of cash balances have been ignored in the literature. Two exceptions, however, are Ronald McKinnon and Chow-Nan Chen. McKinnon argues that under a system of floating exchange rates, in order to facilitate the international flow of commerce, the demand for dollars and all other currencies in which international transactions are conducted would increase. Chen goes even further. He has a model which explicitly incorporates the demand for more than one currency and recognizes that the relative demand depends on the relative opportunity costs. Furthermore, he understands the basic implication of currency substitution, concluding that flexible exchange rates may no longer provide a cushion against foreign shocks. The problem with the Chen model, however, is in the demand for money function. Chen assumes a Cobb-Douglas de-

mand function, which unfortunately constrains the elasticity of substitution to equal one. From the above discussion it should be obvious that a very important empirical question is the precise value of this elasticity. The Chen model therefore cannot be accepted without considerable empirical investigation.

Instead the functional form to be estimated is derived from a procedure similar to one employed by V. Karuppan Chetty. Real balances, denominated in terms of both domestic and foreign currencies, from an individual's cash balance portfolio are combined in a production function for money services. Given the relative efficiencies of domestic and foreign currencies in producing money services (defined by the production function) and the relative opportunity costs of holding different currencies (reflected in the asset constraint), the individual tries to maximize the production of money services.

More specifically, if a *CES* production function is assumed, the level of money services produced by M_d/P_d domestic currency real balances and M_f/P_f foreign currency real balances is

$$(2)\quad \frac{MS}{P_d} = \left[\alpha_1\left(\frac{M_d}{P_d}\right)^{-\rho} + \alpha_2\left(\frac{M_f}{P_f}\right)^{-\rho}\right]^{-(1/\rho)}$$

where

MS = level of money services
M_d, M_f = the domestic currency- and foreign currency-denominated cash balances held
P_d, P_f = domestic and foreign currency price indices
α_1, α_2 = weights reflecting the efficiency of domestic and foreign real balances in producing money services

This production function directly relates the level of real balances to the level of money services. Notice that since real balances in both currencies are in goods units, there is no need for an exchange rate. However, for empirically estimating this relationship, it is desirable to express the production function in terms of nominal cash balances and exchange rates. Defining the exchange rate as $e = P_d/P_f$ from purchasing power parity, and since P_d and P_f are indices, after defining $P_d = 1$, equation (2) becomes

$$(3)\quad MS = (\alpha_1 M_d^{-\rho} + \alpha_2 e M_f^{-\rho})^{-(1/\rho)}$$

The asset constraint for money balances is constructed to reflect two factors: (a) that there is an opportunity cost to holding real balances, and (b) this opportunity cost may differ between the two types of real balances. The overall portfolio of the nonbank private sector of the country is assumed to consist of holdings of all types of real assets, only one of which is money. In constructing that portfolio, desired amounts of each of these assets are determined. Once the demand for each individual asset has been determined, asset constraints for each asset can be constructed. The asset constraint in this paper reflects such an asset demand. It is assumed that in determining the composition of the overall portfolio, the private sector decides to hold M_o real cash balances. These cash balances are then divided between M_d/P_d domestic currency-denominated real balances and M_f/P_f foreign currency-denominated real balances on the basis of the relative cost of holding these different types of balances (reflected in the asset constraint) and their relative efficiencies in providing money services (reflected in the production function).

The asset constraint is of the form

$$(4)\quad \frac{M_o}{P_d} = \frac{M_d}{P_d}(1 + i_d) + \frac{M_f}{P_f}(1 + i_f)$$

where i_d and i_f are the interest rates on domestic and foreign currencies balances, respectively. In terms of nominal balances and exchange rates equation (4) becomes

$$(5)\quad M_o = M_d(1 + i_d) + eM_f(1 + i_f)$$

The asset constraint reflects the fact that M_o is the total money assets that must be held to provide the money services of M_d and eM_f money assets. If for example the money balances are borrowed each period, since it costs $M_d \cdot i_d$ and $eM_f \cdot i_f$ to borrow M_d and

eM_f balances, respectively, a total of $M_d(1 + i_d)$ and $eM_f(1 + i_f)$ money balances must be held in order to pay off the loans at the end of the period.

Maximizing the production function subject to the asset constraint provides the following marginal conditions:

$$(6) \qquad \partial MS/\partial M_d = \lambda(1 + i_d)$$

$$(7) \qquad \partial MS/\partial M_f = \lambda(1 + i_f)$$

$$(8) \qquad M_0 = M_d(1 + i_d) + eM_f(1 + i_f)$$

where λ is the Lagrangean multiplier. Dividing (6) by (7) relates the relative marginal productivities of the two types of balances to their relative prices:

$$(9) \qquad \frac{\alpha_1}{\alpha_2}\left(\frac{M_d}{eM_f}\right)^{-(1+\rho)} = \frac{1 + i_d}{1 + i_f}$$

Taking the logarithm of both sides, rearranging some terms and adding a disturbance term provides the functional form for the estimation:

$$(10) \quad log\frac{M_d}{eM_f} = \frac{1}{1+\rho}\, log\,(\alpha_1/\alpha_2) + \frac{1}{1+\rho}\, log\left(\frac{1 + i_f}{1 + i_d}\right) + u$$

This functional form allows not only the direct estimation of the elasticity of substitution between domestic and foreign currency balances ($\sigma = (1/1 + \rho)$), but it also permits the estimation of the ratio of the coefficients α_1 and α_2. Close substitution between the two types of currency can be reflected in two ways. One way is to have a high elasticity of substitution between the two assets. Another indication of close substitution is similar values of α_1 and α_2. Recall that the α values represent the efficiency of different money assets in providing money services. If foreign and domestic money are perfect substitutes, they should be equal in providing these desired services, and thus the ratio α_1/α_2 should be equal to one. So once the value of σ has been directly estimated as the coefficient of the $log\,(1 + i_f)/(1 + i_d)$ term, the value of α_1/α_2 is computed from the constant term C as

$$(11) \qquad \alpha_1/\alpha_2 = \exp\,(C/\sigma)$$

The model is estimated for Canada using equation (10). Quarterly data on Canadian holdings of *U.S.* dollar (*US$*) and Canadian dollar (*C$*) balances are used as the foreign and domestic currency money balances, respectively. Canadian dollar balances are defined as currency and privately held deposits in Canada. The *U.S.* dollar balances are a sum of *U.S.* dollar liabilities of both *U.S.* and Canadian banks to nonofficial, nonbank Canadians. These *US$* liabilities are then converted to the equivalent *C$* amount using the current end of period exchange rate.

The opportunity cost of *US$* and *C$* balances is measured by the yield on *U.S.* and Canadian Treasury Bills, respectively. It is assumed that the yield on these short-term notes closely reflects the cost of borrowing money. However, since the yields on these securities are reported on a slightly different bases, the *U.S.* Treasury Bill rate is first converted to an equivalent basis to the Canadian Treasury Bill rate.

The results of estimating (10) over the period 1960IV 1975IV are

$$(12) \quad log\,(C\$/US\$) = \underset{(18.0)}{2.56} + \underset{(2.59)}{5.43}\; log\left(\frac{1 + i_{US}}{1 + i_C}\right)$$

$$\bar{R}^2 = 0.78 \qquad F(1.58) = 215.6$$

$$Rho = 0.88 \qquad D.W = 1.44$$

In order to eliminate the presence of first-order autocorrelation in the residuals of the initial ordinary least squares equation, this equation was estimated using a Cochrane-Orcutt procedure. The estimated value of the elasticity of substitution is large at 5.4. It is significantly different from zero at the 99 percent level (two-tailed test), and also significantly different from one at the 95 percent level. Thus the hypotheses that foreign and domestic currencies are nonsubstitutes or that the Cobb-Douglas is the proper money services production function can both be rejected.

The alternative measure of substitution is

TABLE 1—ESTIMATES OF THE ELASTICITY OF SUBSTITUTION DURING SUBPERIODS OF FIXED AND FLOATING EXCHANGE RATES

Subperiod	Exchange Rate Regime	Type of Equation	Constant Term	Elasticity of Substitution	$\bar{R}^2$	D.W.	F	Rho
1960IV-1962II	Floating	OLSQ	2.78 (50.8)	12.8 (2.54)	0.48	1.66	6.47	
1962III-1970II	Fixed	CORC	2.31 (12.7)	2.66 (0.79)	0.78	1.41	107.4	0.9
1970III-1975IV	Floating	CORC	2.79 (16.1)	5.78 (1.83)	0.79	1.27	74.0	0.8

Note: t-statistics in parentheses.
Source. Treasury Bulletin and *Statistical Summary*, Annual Suppl.

to compute the ratio of the efficiency coefficients from the estimated coefficients. This procedure will provide a measure of whether 5.4 is sufficiently close to infinity. If *U.S.* and Canadian dollars are equally efficient in providing monetary services to Canadians, the ratio should equal one.

The initial test on the ratio of the coefficients is to examine the t-statistic on the constant term of the regression. If α_1/α_2 equals one, the logarithm of the ratio will equal zero and the constant term should not be significantly different from zero. However, the significant t-statistic indicates that the value of the ratio is not precisely one. An actual estimate of the value of the ratio is obtained by substituting values from (12) into (11), yielding $\alpha_1/\alpha_2 = 1.60$. From this value it can only be concluded that *U.S.* and Canadian dollars are not perfect substitutes for the entire period.

This estimating procedure is repeated for three subperiods. For the purposes of this analysis it is fortunate that Canada has experienced periods of both fixed and floating rates. The values of the elasticities of substitution under the different exchange rate regimes can now be estimated. Two possible hypotheses concerning these values arise. One hypothesis states that the only reason that the elasticity of substitution was high for the period as a whole was that for a significant subperiod the exchange rate of Canada was fixed. The Bank of Canada was willing during this subperiod to exchange Canadian dollars for *U.S.* dollars, and all the elasticity of substitution was measuring was the substitution on the supply side during this subperiod. This hypothesis would be consistent with high values of elasticities of substitution during fixed rate periods and low values during floating rate periods.

The second hypothesis has just the opposite conclusion. It states that during periods of fixed rates the public does not have to substitute between currencies in private markets since the government is already making currencies perfect substitutes on the supply side. Alternatively, during floating rate periods the public will have to resort to performing all of its substitution through private markets. This hypothesis would be consistent with low or insignificant estimates of the elasticity of substitution during fixed rate periods when the substitution mechanism is not needed, but large estimates during the floating rate periods.

Canada was on floating rates until May 2, 1962, and returned to floating rates on June 1, 1970. The subperiods examined were therefore 1960IV-1962II (floating), 1962III 1970II (fixed), 1970III-1975IV (floating). The results are presented in Table 1. The most striking difference between the subperiods is that in subperiods where the exchange rate was floating the estimate of the elasticity of substitution is larger and statistically significant, while in the subperiod where the exchange rate was fixed, the estimated coefficient is smaller and insignificantly different from zero. For example, in the first subperiod where the exchange rate was floating the estimated value of σ is 12.8, and that value is sig-

nificantly different from both zero and one at the 95 percent level (one-tailed test). Similarly, in the final subperiod where the exchange rate was again floating, the estimated value of σ is 5.8, and that value is significantly different from zero at the 95 percent level and from one at the 90 percent level. In contrast the estimated value of σ for the subperiod where the exchange rate was fixed is less than half the value in any other subperiod or the period as a whole. In addition, the coefficient is not significantly different from zero at even the 90 percent level. Not surprisingly the ratio of the efficiency parameters exhibit a similar pattern with estimates of 1.2 and 1.6 in the first and second floating rate periods and 2.4 in the fixed rate period.

The results from analyzing the subperiods are therefore consistent with the second hypothesis and not the first. The large significant elasticity of substitution for the period as a whole does not seem to be the result of substitution during the fixed rate subperiod, but rather during the floating rate subperiods. The concept of substitution between *U.S.* and Canadian dollars by private Canadians appears to be statistically valid when the Canadian government is not performing that service for them. The results are even more impressive when one considers that the above data have not even included holdings of *U.S.* dollar-denominated Eurodollars, whose elasticity of substitution with respect to Canadian dollars could quite possibly be even higher.

IV. Summary and Implications

In this paper it has been argued that there exists a group of individuals within a country who diversify their real cash balance holdings between domestic and foreign currency-denominated balances. These diversified portfolios imply that monetary policy will produce changes in the interest rate that induce offsetting money flows even under perfectly flexible exchange rates. The importance of these offsetting flows will be directly proportional to the degree of substitution between currencies. In the case where currencies are perfect substitutes monetary independence is impossible, even where central banks do not intervene in the foreign exchange markets.

Thus the argument that flexible exchange rates imply monetary independence is brought into question. Only to the extent that individuals do not substitute between currencies will the argument be valid. As the empirical tests show, the elasticity of substitution in at least one major country is quite high. The proper model for analyzing monetary policy may therefore be one of monetary dependence, not monetary independence, even when perfectly flexible exchange rates are assumed.

REFERENCES

C.-N. Chen, "Diversified Currency Holdings and Flexible Exchange Rates," *Quart. J. Econ.*, Feb. 1973, *87*, 96–111.

V. K. Chetty, "On Measuring the Nearness of Near-Moneys," *Amer. Econ. Rev.*, June 1969, *59*, 270–81.

Milton Friedman, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago 1953.

———, *The Optimum Quantity of Money and Other Essays*, Chicago 1969.

Ronald McKinnon, *Private and Official International Money: The Case for the Dollar*, in *Essays in International Finance*, Princeton Univ. No. 84, Apr. 1969.

Robert A. Mundell, *International Economics*, New York 1968.

Bank of Canada, *Statistical Summary*, Annual Suppl., Ottawa 1960–69.

———, *Review*, Ottawa 1971–76.

U.S. Treasury Department, *Treasury Bull.*, Washington 1960–76.

The Pure Theory of the Muggery

By PHILIP A. NEHER*

Dear Old Friend: As you have doubtless observed, it has been many a week now since I last stepped out of the shadows, put the barrel to your ribs and divested you of cash, trousers, and credit cards. You have probably asked yourself, "Where is my faithful old stickup man, John? Doesn't he like me any more?"—The truth is, sir, that I'm terrified you'll take it amiss and move out of New York if I overwork my welcome.

John

[Quoted in Russell Baker's column, *New York Times*, December 9, 1975]

The muggery is a geographic place where muggers and muggees transfer wealth from the latter to the former. Such transfers involve the allocation of time and money, so it seems natural to view mugging as an economic activity with the agents on both sides maximizing subject to constraints. By taking this approach, I follow Thomas C. Shilling, Gary Becker, Isaac Ehrlich, and others who have brought traditional tools of economic analysis to bear on the understanding of criminal activity.

There is a related literature which is also relevant. I observe that the muggery has two features in common with the open sea fishery.

(1) Muggers and muggees stand in a fundamental *predator-prey relationship* to each other just as fisherman and fish.

(2) As prey, muggees are a *common property resource* from the point of view of an individual mugger.[1] Over-mugging can occur for the same reasons as over-fishing.

I shall explore this analogy in subsequent sections. First, a dynamic model of an uncontrolled muggery is developed. It generates a rich variety of time-path solutions which may help explain the diverse experience of different urban areas in generating mugging activity. Second, I employ modern capital theory to demonstrate an incentive to organize and control the muggery. Finally, I suggest some extensions of the analysis to other forms of economic activity. I conclude by asserting that allocative inefficiencies associated with common property may be widespread and important in enterprise economies.

I. The Free Entry Muggery

To set the scene, consider an urban area where people normally walk the streets in possession of mobile wealth: money, watches, jewelry, and the like. These people are potential muggees, henceforth b's. The fact that b's voluntarily frequent the muggery is evidence that they derive benefits from doing so: travelling to and from work, shopping, seeking entertainment, and so on. On the other hand, they are deterred from entry by their apprehension of being mugged and relieved of wealth thereby.

Muggers, henceforth a's, are attracted to the muggery by the prospect of potential wealth transfers in their favor. But to engage in mugging entails the cost of foregone leisure and of earnings in other activities. These costs arise because the act of mugging takes time to search, hit, and avoid capture, and because of possible arrest and

*University of British Columbia. I wish to thank G. C. Archibald, C. Blackorby, D. J Donaldson, B C Eaton, S Q. Lemche, K. Nagatani, and A D Scott for their comments on an earlier version Usual disclaimers. Financial support of the Canada Council is gratefully acknowledged

[1]The classic articles on fishery economics are by Anthony Scott and H Scott Gordon For a recent and accessible bibliography, see Colin Clark and Gordon Munro

incarceration.

In short, both a's and b's incur both benefits and costs from participation in the muggery.[2] The difference I shall refer to as "profit."

Following Vernon Smith I assume that a's and b's will increase in numbers if the respective benefits of participation in the muggery exceed the costs. That is, "firms" will enter an "industry" in response to industry profits.

Formally, entry and exit is described by a system of differential equations[3]

$$Da = F(a, b) \tag{1a}$$

$$Db = G(a, b) \tag{1b}$$

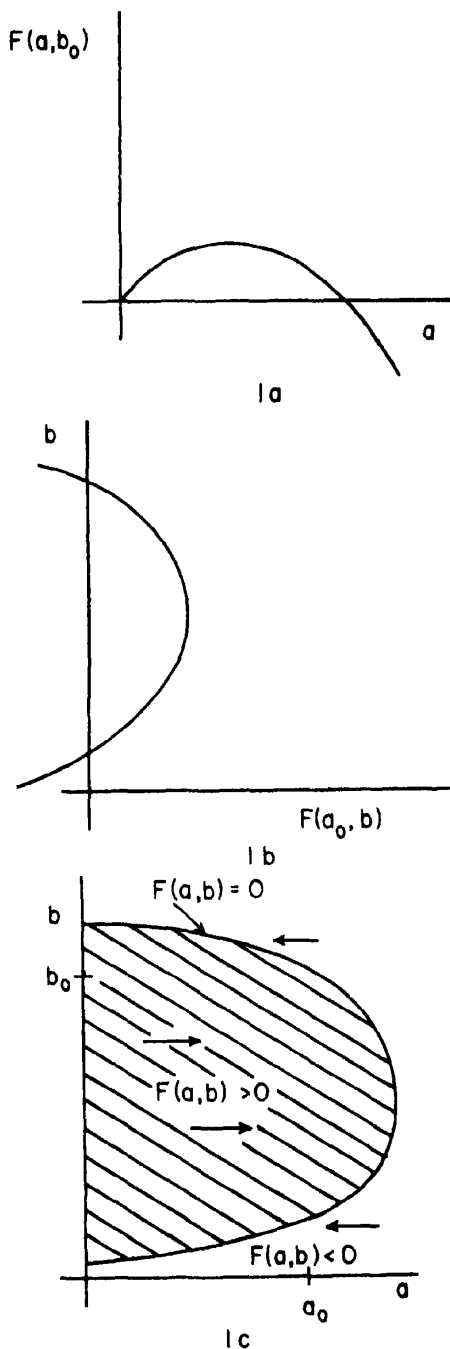
where D denotes the first time derivative. The *flow* of a and b in and out of the muggery depends on the *stocks* of both a and b. I now specify the nature of those dependencies in terms of the first- and second-order partial derivatives of $F(\ )$ and $G(\ )$.

A. *Muggers*

Set the number of b's at some arbitrary number, $b = b_0 \geq 0$. Then conduct a conceptual experiment. How will total mugging profits vary, and thus the entry of a's vary, as the number of a's increase from zero? If a is zero, then total revenue, costs, and profits are zero. Adding a's, more of them mean more successful hits and more revenue, along with rising total costs. But it is reasonable to assume that marginal revenue falls as successive a's prove less adept at mugging. Likewise, one would expect marginal opportunity costs to rise as successive a's abandon more valuable leisure and higher paying occupations elsewhere.

FIGURE 1

[2]Throughout this paper, I shall assume that a's and b's are two distinct groups. Cross overs do not occur I do this to avoid facing up to difficult problems of occupational choice. I shall also assume that a's do not mug a's.

[3]The functions $F(\ )$ and $G(\ )$ are really profit functions. But through an appropriate choice of profit measurement units, I can use them to denote the speed of entry and exit as well.

Taken together, falling productivity and rising costs would seem to ensure diminishing marginal profits as more of a is added. A cross section of the profit functions, with $b = b_0$, appears in Figure 1a to illustrate these considerations.

Next, set the number of a's at some arbitrary number, $a = a_0 \geq 0$, and vary the number of b's. For small numbers of b's, profits are negative because the opportunity cost of search looms large compared with revenues from the occasional hit. On the other hand, large numbers of b's provide "safety in numbers" through "help thy neighbor" effects and through organized law enforcement (police) efforts which can take advantage of economies of scale. Taken together, these effects would seem to ensure negative profits for high b densities. Figure 1b illustrates these considerations.

Information contained in Figures 1a and 1b is depicted in Figure 1c. The profit function appears to be shaped like a truncated "dome" and, in the usual way, one can plot iso-profit lines by passing level surfaces through the dome parallel to the ab plane, intersecting the Da image. Everywhere within the shaded area $Da = F(a,b) > 0$. Outside, $Da = F(a,b) < 0$. In between these areas, $Da = F(a,b) = 0$. These are the zero profit and zero growth combinations of a and b. Arrows denote the movement of a elsewhere.

B. *Muggees*

Set $b = b_0 \geq 0$ and vary a. How will muggee profits, and thus the growth of b's, be affected? Taking benefits to be unrelated to the number of a's, profits will fall as more a's impose additional costs on the b's. This is illustrated in Figure 2a.

Next, set $a = a_0 \geq 0$ and vary b. As b's increase, total benefits can be expected to increase, at least up to some point where "crowding effects" reduce local environmental quality. Total costs will increase as the given number of a's make more hits. Taking these considerations together, it is difficult to know how the number of b's af-

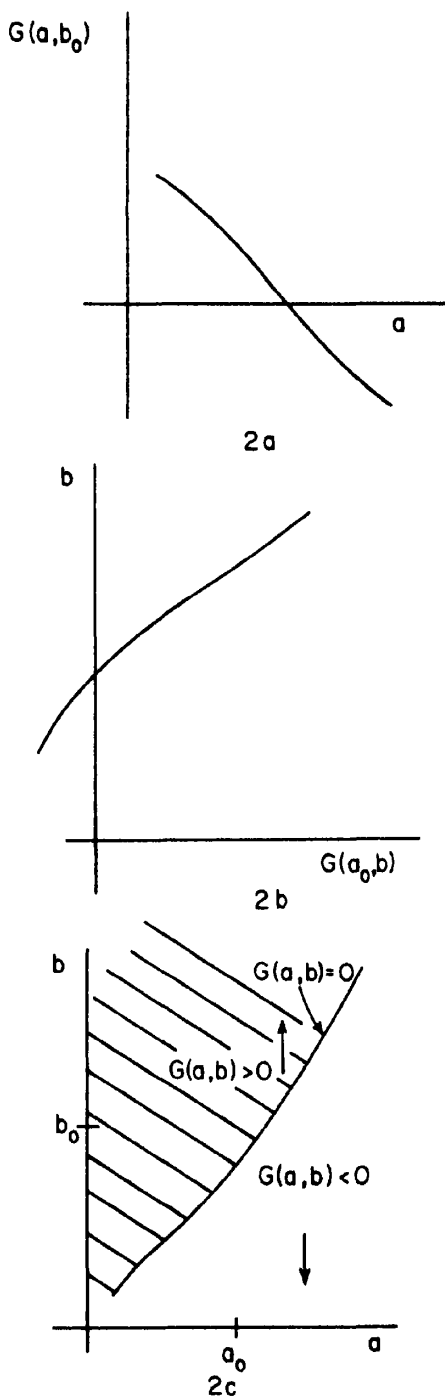



FIGURE 2

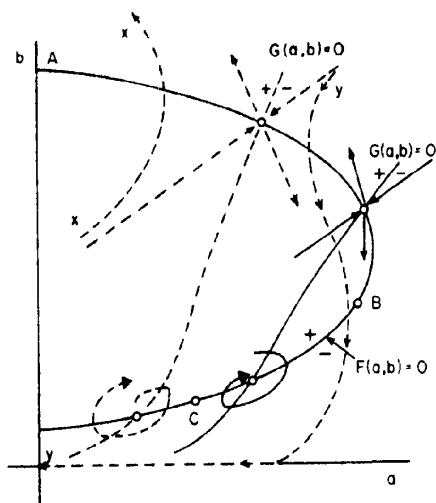



FIGURE 3

fects the growth of b's. To preserve simplicity, I shall assume a monotonic positive relationship: more b's encourage the growth of b's. This relation is illustrated in Figure 2b. Casual empiricism suggests that it is appropriate for a wide range of urban experience which I want to investigate. Figure 2c combines information from Figures 2a and 2b.

The dynamics of the muggery can now be investigated. Figures 1c and 2c are superimposed in Figure 3 where the motions of a and b are depicted by arrows and specimen trajectories.

Evidently, a variety of qualitatively different experiences are possible. Look first at possible singular points (where $Da = Db = 0$). Two possible $Db = 0$ lines are depicted: one is broken, the other is solid. Between points A and B in Figure 3 the slope of the $Db = 0$ line exceeds the slope of the $Da = 0$ line, the roots of the system are real, so singularities are saddlepoints. Moving clockwise past point B, the roots become imaginary. At first $F_a + G_b$ may be negative, due to F_a being more negative than G_b is positive, so singularities are stable foci. Continuing clockwise, F_a is becoming larger. At point C, $F_a + G_b$ equals zero and the singularity is a center. Past point C, $F_a + G_b$ is positive but the roots remain imaginary, giving rise to unstable foci.

Consider next possible initial conditions. These, along with information on the singularities, permit construction of a variety of trajectories which illustrate different histories of the muggery. For the most part, I leave this task to the reader, but certain classes of outcomes seem worthy of comment.

A. A sufficiently high density of b's relative to a's will eventually result in all the a's being driven out. Initial conditions lying above stable arms between A and B are sufficient for this to occur.[4] The broken xx curve is an example. Initially "safe" areas become even safer. I call this the "Scarsdale effect." By extension, one can imagine areas which are safe, or not, depending on the time of day. The theatre district in New York City is reported relatively free of mugging when theatre goers, presumably b's, are dense.

B. Suppose $G(a,0) < 0$. Then any trajectory intercepting the a axis will eventually result in the elimination of a's as well as b's. The broken yy curve is an example. I call this the "Dallas effect." Virtually no one walks the streets so it does not pay a's to search out the occasional b who, as a consequence, is perfectly safe.

C. Oscillations about a stable focus can culminate in an equilibrium muggery with relatively high crime rates. The a's and the b's have reached an accommodation with each other. No person perceives it to his advantage to move into or out of the muggery.

II. The Controlled Muggery

Even the casual mugger could not fail to be impressed with the fact that his own success is affected by the activities of other muggers. If the muggery is truly competi-

[4] Some exogenous constraint, not imposed here, is clearly required to put an upper bound on b density.

tive, external effects imposed by any one mugger on any one other are, of course, small.[5] But taken together, competitive muggers will reduce their total profits to zero. This is the classical rent dissipation solution observed in the uncontrolled fishery by Scott and by Gordon.

Each competitive mugger will regard potential muggees as a common property resource. If property rights are not assigned and enforced, potential rents will be dissipated. The tragedy of the commons can characterize the muggery as well as the fishery.

Consider a single act of mugging. The mugger knows that the act reduces his own chances of future success since the muggee is taken, at least temporarily, out of the prey population and, in the long run, can be expected to take more permanent evasive action. Moreover, the act is a signal to other potential muggees that they too could be victims.[6] These effects will be taken into account by the individual mugger insofar as they bear on him. But the individual will not care that his own activity reduces the success that could have been enjoyed by others in the future. The consequent "inefficiency" is an incentive to organize the muggery: to internalize these external effects by establishing property rights over the muggery.

Assume that the muggery manager wishes to maximize the discounted sum of profits (wealth transfers minus costs) over time.

$$\max J = \int_0^T F(a,b)e^{-rt}\,dt \tag{2}$$

In this formulation, $F(\ )$ is drawn from (1a), $T \leq +\infty$ is the manager's time horizon, and r is his discount rate.[7] To solve the maximizing problem, the manager will be concerned not only with the current flow of profits $F(\ )$, but also with the imputed value at the shadow price (p) of investments made in the stock of muggees (b). These considerations are captured in the present value Hamiltonian which represents the value of "consumption" plus "investment."

$$H = E^{-rt}[F(a,b) + p \cdot Db] \tag{3}$$

$$Db = G(a,b) \tag{1b}$$

Equation (1b) is now recognized as the dynamic constraint facing the manager. He knows and cares that $G_a < 0$ and $G_b > 0$, from Figure 2. It is precisely these effects which are collectively disregarded by free entrants, but which are crucial to the manager. One suspects, and it turns out to be the case, that the manager will assign a positive shadow price p to Db, while a free entrant will not. There will be a conflict of interest between managed muggers and interlopers. The former will enjoy a rent which the latter would like to capture.

To solve the maximizing problem posed by (2), the manager should first arrange to manage that which he can. In this case, I assume the manager can costlessly control the number of a's. He should adjust a at every point in time to maximize (3). Assuming a regular interior maximum, he adjusts a so that $H_a = 0$, or

$$F_a = p \cdot (-G_a) \tag{4}$$

The addition to current profits by taking on another a is F_a. But by doing so, some b's escape, as denoted by G_a. Each b has a value p. Thus (4) is the usual hiring condition: current gains and losses should balance at the margin.

Having discovered in (4) that p matters,

[5] The prefatory quote suggests quite the opposite. It is clear that John thinks of himself as having some monopoly over the mugging of Russell Baker. John is, I presume, a competitive mugger but, apparently, an imperfect one. John's sense of husbandry is imputed to a muggery manager in this section.

[6] The analogy to the fishery breaks down here. I do not believe that fish learn from the experience of other fish.

[7] It is not clear how T and r are determined. If the muggers are organized as a democratic worker-managed firm, then both T and r will depend on the demographic characteristics of the workers. See Irving Fisher and the author.

the manager would like to know how to assign its value. Recognizing that b is capital, and p is its price, the manager simply applies the zero net profit condition $D(e^{-rt}p) = -H_b$ or

$$(5) \qquad Dp + F_b + p\cdot G_b - r\cdot p = 0$$

Taken together, the first three terms are the marginal net benefits of having muggees, a form of capital, in the muggery. The first represents capital gains. The second term is the value of the marginal contribution of b's to the marginal profits reaped by the a's. The third is a physical appreciation (negative depreciation) term. Since G_b is positive, $p\cdot G_b$ is the value contribution to a's profit of having another b to mug. The fourth term is the marginal rental price of muggees.

In terms of conventional capital theory, (5) states the optimal usage of b requires that the value of capital gains, *plus* the value of capital's marginal production, *minus* physical depreciation, *minus* rental charges must equal zero. The final condition is the requirement that the dynamic constraint (1b) be observed.

Equations (4), (5), and (1b) are in the form

$$(4') \qquad 0 = f(p,a,b)$$

$$(5') \qquad Dp = g(p,a,b)$$

$$(1b') \qquad Db = h(p,a,b)$$

By solving for a in (4') (by virtue of the implicit function theorem), and then substituting its value into (5') and (1b'), one obtains

$$(6a) \qquad Dp = j(p,b)$$

$$(6b) \qquad Db = k(p,b)$$

Equations (6) are autonomous of time, so that the well-developed theory of plane autonomous systems can be utilized to understand the motion of the controlled muggery. The addition of appropriate initial and terminal (or transversality) conditions could close the system.

Instead of exploring the implications of (6) in detail, I want to focus on properties of the stationary state, where capital gains have withered away ($Dp = 0$) and the muggee population has stabilized ($Db = 0$). In that state, (5) can be written

$$(7) \qquad \frac{F_b + G_b\cdot p}{r} = p$$

The shadow price of an additional muggee is the capitalized value of its net marginal benefits.

It is now easy to expose the economic meaning of overmugging if a's have free access to b's. The uncontrolled a has but small concern for the consequences of his activities on the population of b's. From his point of view, a foregone hit may be captured by somebody else.[8] He thus has little incentive to husband the b resource by placing a shadow price p on its increase. He values the future of b's not at all.

In terms of (7), this lack of concern can be captured by an infinitely large discount rate. Letting r get indefinitely large, p must get indefinitely small. The free access muggery, in this sense, will be simulated by a controlled access muggery whose manager cares not about the future.[9]

But if the a's can organize and establish property rights over the b's, the muggery can support more of both. Moreover, the greater is the sense of husbandry that the a's feel for the b's (the lower is r), the more each will flourish.

Contemplating these results, should society prefer competitive, free entry, mugging over the same activity organized by a far-sighted manager? The question is complex and I hesitate to comment. But one judgement is easy to make. If one places a sufficiently high value on having viable urban neighborhoods, organized mugging is to be preferred. Well-organized and managed muggers will not drive their prey to extinction and legitimate economic activity will survive. If this is desired, police

[8] This attitude can motivate "forestalling" behavior in the exploitation of many natural resources where property rights over them are not assigned and enforced. One forestalls his competitors by getting there first.

[9] See Stephen Chung and Clark. My results are consistent with theirs.

might actually encourage "neighborhood gangs" by reducing the transactions costs of organizing (by subsidizing "youth centers," for example) and by helping to erect and maintain barriers to entry.

But if the muggery is controlled, there are not only more muggees, but also more muggers. Acts of muggery will be more frequent. These results are consistent with those already well established in the fishery literature. A controlled fishery supports more fish and fishermen, and more fish are captured. But the social value of having fished a fish is less ambiguous than the value of having mugged a muggee. People do not assign equal welfare weights to people and fish. However, muggers and muggees are both people, and while one might well be agnostic about wealth transfers between them, the concomitant potential violence is surely to be deplored. Moreover, organized gangs are typically multiproduct firms. One might welcome control of some of their activities while deploring it in others.

III. Extensions

The muggery is naturally understood as an institution where predator-prey relationships are important, and where questions of allocative efficiency associated with common property can arise. But these characteristics extend to other forms of economic activity: legal and illegal, market and nonmarket. I want to suggest that many of these can be better understood in the predator-prey context.

Put another way, in what circumstances would one *not* want to deploy our understanding of predator-prey relationships and possibly associated common property difficulties, to analyze economic activity? I think that the only sure ground is to be found in the Kenneth Arrow and Gerard Debreu world of perfect and complete markets, which are organized by an unpaid auctioneer. In that world, there are no market failures. Property rights are assigned and (costlessly) enforced. In other circumstances, the predator-prey model applies, with a force depending on the degree of market failure. Consider some examples.

A. *Computer Crime*

Mr. Smith is a clever but dishonest computer programmer. He discovers a method to get his Book-of-the-Week Club to write him checks by punching appropriate holes in his weekly IBM card. If Mr. Smith, and his method, are found out, then similar crime becomes more difficult not only for Mr. Smith, but also for others, as his book club takes evasive action. As a predator, Smith makes criminal activity less profitable for other potential predators. He has no incentive to take this external effect into account. Other forms of illegal activity which involve true victims are characterized by similar externalities. An illegal act will alert potential victims who consequently undertake to protect themselves not only from the original predator, but also from persons with similar intent.

B. *Wolf Cub Bottle Collections*

In this case the Cubs are the predators and households having bottles to collect are the prey. Free entry bottle collecting would entail multiple canvassing and consequent waste of Cub time. The analogue to the fishery goes through intact: a collected bottle is a fished fish. Optimal bottle collecting requires centralized control.

C. *The Household Encyclopedia Market*

The second-hand market for home encyclopedias is notably bad or, to say the same thing, the rental market for encyclopedia services is not well-developed. If a household buys an encyclopedia from one salesman, it has been fished out from the point of view of other salesmen. This is a "thinning" effect which is recognized in the fishery as a technological externality. To discover it in a conventional market may seem a bit odd. But just as in the computer crime and in the bottle collecting, the analogy to the fishery goes through. Again,

efficiency seems to call for centralized control. But that would have the unfortunate side effect of conferring monopoly power to a seller in a market where one would like consumer sovereignty over product choice to prevail. Evidently, one cannot have both canvassing efficiency and a full consumer choice at the same time.

The problem is rooted in the failure of market participants (in this case, buyers) to generate information. The market would work better if each household, having purchased an encyclopedia, would advertise its market position. The trouble is that information is, in large measure, a public good. The household has no way to internalize the benefits from its production. It cannot claim and enforce a property right over the social value of that right. One expects, then, that market signals to the effect "I have an encyclopedia and don't want another" will not be generated at a socially optimal level. This does not mean that one would not observe, say, "No Peddlers" signs. But one would expect them to appear at a frequency which is less than socially optimal.

V. Conclusions

The concept of property rights has become central to the understanding of economic behavior. Its modern elaboration is due to Harold Demsetz, Armen Alchian, James Buchanan and Gordon Tullock, John Dales, and others cited by Eirik Furubotn and Svetozar Pejovich in their review article. The concept is crucial in the applied field of natural resource economics, particularly fishery economics. So it is not surprising that the applied work of Scott, Gordon, and others should have paralleled these related developments.

I am suggesting a close analogy between the fishery and the muggery. The muggery, in turn, is a paradigm of criminal and other economic activities where predator-prey and common property relationships are particularly important, and where the prey has the property of being capital from the predator's point of view. I hesitate to pronounce on the frequency of these situations, but I am confident in asserting that they occur more often, and are more important, than most of us have recognized.

REFERENCES

A. Alchian, "Corporate Management and Property Rights," in Henry Manne, ed., *Economic Policy and the Regulation of Corporate Securities*, Washington 1969, 337–60.

K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev. Econ Stud.*, Apr. 1964, *31*, 91–96.

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *78*, 526–36.

James Buchanan and Gordon Tullock, *The Calculus of Consent*, Ann Arbor 1962.

S. N. S. Chung, "Economics of Fisheries Management A Symposium," Instit. Animal Resource Ecology, Univ. British Columbia, Vancouver 1970

C. W. Clark, "Profit Maximization and the Extinction of Animal Species," *J Polit Econ.*, July/Aug. 1973, *81*, 950–61.

——— **and G. R. Munro**, "The Economics of Fishing and Modern Capital Theory: A Simplified Approach," *J. Environ Econ Manage.*, 1975, *2*, 92–106.

John H. Dales, *Pollution, Property and Prices*, Toronto 1968.

Gerard Debreu, *Theory of Value*, New York 1959.

H. Demsetz, "Toward a Theory of Property Rights," *Amer Econ Rev. Proc.*, May 1967, *57*, 347–73.

I. Ehrich, "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation," *J. Polit Econ.*, May/June 1973, *81*, 521–95.

Irving Fisher, *The Theory of Interest*, reprinted New York 1965.

E. Furubotn and S. Pejovich, "Property Rights and Economic Theory: A Survey of Recent Literature," *J. Econ. Lit.*, Dec. 1972, *10*, 1137–62.

H. S. Gordon, "The Economic Theory of a Common Property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124-42.

P. A. Neher, "Democratic Exploitation of a Replenishable Resource," *J. Publ. Econ.*, Apr./May 1976, *5*, 361-71.

A. D. Scott, "The Fishery: The Objectives of Sole Ownership," *J. Polit. Econ.*, Apr. 1955, *63*, 116-24.

T. C. Shilling, "Economic Analysis of Organized Crime," in President's Commission on Law Enforcement and Administration of Justice, *Organized Crime*, Washington 1967.

V. L. Smith, "Economics of Production from Natural Resources," *Amer. Econ. Rev.*, June 1968, *58*, 409-31.

An Econometric Definition of the Inflation-Unemployment Tradeoff

By GREGORY C. CHOW AND SHARON BERNSTEIN MEGDAL*

In the coming year 1979, is it possible to achieve a 5 percent unemployment rate and keep annual inflation down to 4 percent? To raise this question in terms of a particular econometric model, we ask whether there exist values of the policy instruments which will give rise to solutions of 5 and 4 percent, respectively, for unemployment and inflation. What is the most favorable tradeoff relationship between inflation and unemployment implicit in an econometric model of a national economy? In this paper, we wish to point out that for many econometric models actually in use, the tradeoff relationship is not rigid, but can be shifted toward the origin (but usually not all the way to the origin!) by suitable government policies. Accordingly, we suggest that the tradeoff relationship implicit in an econometric model be defined as the set of points in the unemployment-inflation diagram which cannot be dominated. We will explain the circumstances under which there exists such a southwestern boundary for the points depicting the unemployment-inflation combinations that are achievable according to a given model. We will propose a systematic way to locate points on this boundary and demonstrate that our algorithm works.

Stimulated by and based upon A. W. Phillips' original paper on the relation between unemployment and the rate of change of money wage rates, numerous studies have appeared to refine, respecify, and estimate structural equations explaining the rates of change in the wage rates, the price level, unemployment and related variables. It soon became apparent that these studies, though useful, may not be sufficient for ascertaining the tradeoff relationship between unemployment and inflation. If unemployment and inflation are viewed as two of the many endogenous variables which are jointly determined by a system of simultaneous econometric equations, their relationship has to be derived by solving a whole system using alternative values for the policy variables subject to government control. The approach of deriving the unemployment-inflation tradeoff by varying the policy variables and solving for these two endogenous variables in an econometric model has been adopted by Leonall Andersen and Keith Carlson, Albert Hirsch, George de Menil and Jared Enzler, Saul Hymans, and Ronald Bodkin, among others.

Since, as we shall explain, the relationship between unemployment and inflation implicit in an econometric model is usually not rigid, an arbitrary choice of policy instruments in the above simulation approach will ordinarily produce combinations of these two variables which can be improved upon. We therefore suggest that the policy variables be chosen optimally, rather than arbitrarily, in order to derive the most favorable tradeoff relationship. By choosing two alternative paths for the policy variables arbitrarily, one more expansionary than the other, one would expect to obtain two solutions for unemployment and inflation from an econometric model. However, by a more judicious choice of the policy variables, it may be possible to improve on both of these results. In general, the unemployment-inflation combinations resulting from varying the values of the policy variables in an econometric model would be a scatter and would not all fall on one

*Princeton University. We would like to thank Saul Hymans and Harold Shapiro for generously supplying us information concerning the Michigan Quarterly Econometric Model; Keith Carlson for gracious help with the St. Louis Model; Alan S. Blinder, George H. Borts, and a referee for useful comments; and the National Science Foundation for financial support.

rigid curve. Thus a unique tradeoff relationship may not be obtainable simply by trying out different values for the policy variables. Some form of optimization is required to derive the best possible tradeoff relationship.

In Section I of this paper, we will first point out the various possibilities for unemployment and inflation implicit in a static econometric model consisting of a set of simultaneous equations and propose a method to derive from the model the best tradeoff curve. Section II generalizes the discussion to the dynamic case. Section III applies our approach to derive the best inflation-unemployment tradeoff from the St. Louis Model, and Section IV applies the same for the Michigan Quarterly Econometric Model. Section V contains some concluding remarks.[1]

I. Inflation-Unemployment Possibilities in a Static Model

Given a set of simultaneous equations determining a vector $y' = (y_1, y_2, \ldots, y_p)$ of endogenous variables by a vector $x' = (x_1, \ldots, x_q)$ of policy instruments, and given a set of exogenous variables not subject to government control which will be treated as fixed, we ask what combinations of unemployment y_1 and inflation y_2 are possible and how one can trace out the best possible combinations.

For our purpose, the set of possible solutions for y_1 and y_2 can conveniently be classified into three categories. The first is a rigid relation, the points all falling on one curve in the y_1-y_2 diagram. The second is a semirigid relation, the set of possible points forming an area in the y_1-y_2 plane which has a southwestern boundary. The second case is considered most important and it is the southwestern boundary which we would like to ascertain as the best possible tradeoff relationship. The third is the least rigid, the set of possible points not being bounded by a southwestern boundary. As a special case of the third category, we may have the set of possible points covering the entire y_1-y_2 plane. This is a mathematical possibility, but the model involved would not be economically meaningful because y_1 cannot be negative.

The first possibility, namely a rigid tradeoff curve, occurs when there is a structural equation explaining y_2 by y_1 and some exogenous variables, for example, $y_2 = -1.2y_1 + z$. The variable z in this equation may incorporate exogenous variables and other variables as long as these variables cannot be influenced, directly or indirectly, by the policy instruments. Otherwise the relation between y_2 and y_1 can be shifted. The set z of variables in this structural equation may consist entirely of exogenous variables not subject to government control, or of some endogenous variables which are determined completely by uncontrollable exogenous variables. The equation relating y_2, y_1 and the other variables so specified might not itself be a structural equation, but the result of combining several structural equations. To illustrate, let y_3 be the rate of change in the wage rate. Assume a wage Phillips curve relating y_3 to y_1, y_2 and possibly some exogenous variables not subject to government control. Assume also a price Phillips curve relating y_2 to y_1, y_3 and possibly some exogenous variables. Eliminating y_3 from these two structural equations would yield a rigid tradeoff relationship between y_1 and y_2. In terms of the reduced-form equations determining y_1 and y_2 by $x_1, \ldots, x_q$, the $2 \times q$ matrix of partial derivatives of y_1 and y_2 with respect to the q x's would be of rank 1, so that starting from a given point, when a small change in y_2 results from whatever changes in the x's, y_1 will be changed proportionally in the opposite direction.

If any of the other variables in an equation relating y_1 and y_2 (which may be itself a structural equation or, more likely, the result of combining several structural equa-

[1]In this paper, our main purpose is to propose a definition of the best inflation-unemployment tradeoff and a method of deriving the relationship from an econometric model. We are not concerned with the actual shape of the price Phillips curve, and therefore, would avoid discussion of whether the long-run Phillips curve is nearly vertical.

tions) can be influenced directly or indirectly by the policy variables, the relation between y_1 and y_2 will no longer be rigid. The possible combinations of y_1 and y_2 obtained by varying the x's will form an area in the y_1-y_2 plane. The $2 \times q$ matrix of partial derivatives of y_1 and y_2 with respect to the qx's via the reduced form will have rank 2 for many values of y and x. This gives rise to the second or the third category of possible solutions for y_1 and y_2. In general however, the possible combinations do not take up the entire y_1-y_2 plane as they would in the special case of the third category of our classification. If the equation relating y_2 to y_1 is linear, such as $y_2 = -1.2y_1 + z$, and z is a linear function of policy instruments which can take any positive or negative values, the tradeoff curve can then be shifted at will. On the other hand, a non-linear structural equation explaining the rate of unemployment y_1 may rule out negative values for y_1. For example, if $\log y_1$ but not y_1 appears in the model, y_1 cannot be negative. Furthermore, a southwestern boundary may exist for the possible combinations of y_1 and y_2. This can occur when the values of the x's are bounded (such as the tax rates, money supply, and government expenditures taking only nonnegative values). It can also occur because, in the equation relating y_1 and y_2, the other endogenous variables which can be influenced by government policies are bounded by their own nonlinearities or by the boundedness of the government instruments themselves. Econometric models belonging to our second category, those with a southwestern boundary for the possible y_1-y_2 combinations, appear to be economically reasonable. The first category would rule out the possibility of any bad policies which can make both inflation and unemployment worse. The third category would imply that one can achieve any desired inflation-employment combination as one pleases.[2]

[2] In Chow (1976), it was found that by manipulating the policy instruments, one can reach any desired combination of the level of employment and the general price index according to the Klein-Goldberger model of the U.S. economy.

If a southwestern boundary exists for the possible y_1-y_2 combinations, one can obtain points on this boundary by systematically varying the parameters k_1, k_2, a_1, and a_2 in a quadratic loss function

$$w(y_1,y_2) = k_1(y_1 - a_1)^2 + k_2(y_2 - a_2)^2 \tag{1}$$

and minimizing this function subject to the constraint of the econometric model. To see that one point in the boundary will result from such a minimization, let $k_1 = k_2 = 1$ and $a_1 = a_2 = 0$. The points of equal loss will form a circle with center in the origin, and circles closer to the origin will have smaller losses. The minimum occurs when the smallest circle is tangential to the boundary of the possible y_1-y_2 points constrained by the econometric model. Assume that, in the first quadrant of the y_1-y_2 plane, a southwestern boundary exists, which means that the slope of the boundary is negative (or at least nonpositive). Since the slope of the circle in the first quadrant is also nonpositive, obtaining the smallest circle satisfying the constraint will mean that the y_1-y_2 point is on the boundary—if it were not, one could use a smaller circle satisfying the constraint and reducing the loss. To obtain another point on the boundary, one could change the ratio of k_1 to k_2, letting $k_1 = 100$ and $k_2 = 1$, say. The points of equal loss would be an ellipse which is elongated vertically. In sacrificing one unit of unemployment, one requires a greater reduction in inflation than before; the slope of the ellipse in the first quadrant of the y_1-y_2 plane is steeper than before. The new minimum will yield a smaller unemployment and a higher inflation rate. We can drop the above assumption that the southwestern boundary lies in the first quadrant. If it were in the fourth quadrant, the above analysis would apply by placing the center (a_1, a_2) of the ellipse below and to the left of the boundary.

Although minimization of (1) with $k_1 = k_2 = 1$ and $a_1 = a_2 = 0$ will yield a point on the best tradeoff curve in the first quadrant if it exists, one cannot anticipate the resulting value for either y_1 or y_2. To answer the

question, what is the lowest inflation rate y_2 for a given unemployment rate $y_1 = 5$ (percent), one may set $k_1 = 1000$, $k_2 = 1$, $a_1 = 5$, and $a_2 = 0$. The points of equal loss form a highly vertically elongated ellipse with (5, 0) as center. Minimization yields an unemployment rate close to 5 percent and the corresponding lowest inflation rate, since it chooses the smallest vertical ellipse centering in (5, 0) which still satisfies the constraint of the econometric model. By the same argument, replacing $a_2 = 0$ by $a_2 = 2$ in the above minimization would also work provided that the lowest inflation rate for 5 percent unemployment is above 2 percent.

In this section, we have classified the possible solutions for unemployment and inflation from a static econometric model, defined the best tradeoff relationship implicit in the model, and suggested a method for tracing out this relationship.

II. Unemployment-Inflation Tradeoff Possibilities in the Dynamic Case

It is important to generalize our discussion to the dynamic case because econometric models are dynamic, and economists are interested in the best tradeoff relationships between unemployment and inflation through time. In the dynamic setting, we have to consider the three-dimensional space with unemployment y_1, inflation y_2, and time t as the axes. The three categories of tradeoff possibilities implicit in an econometric model will be discussed for the very short run, the intermediate run, and the very long run.

By the very short run, we mean one quarter if the econometric model is a quarterly model. The dynamic model consists of y_t, y_{t-1}, and x_t as variables, the uncontrollable exogenous variables being considered given as before. Endogenous variables lagged more than one period and lagged policy instruments can be eliminated from any dynamic model by introducing suitable identities, as explained in Chow (1975). For the one quarter immediately ahead, all lagged variables y_{t-1} are given, and the analysis reduces to the static case. The discussion of Section I applies entirely to the tradeoff possibilities between y_{1t} and y_{2t} given all lagged variables.

The intermediate run requires some discussion. When the time interval of interest is from period 1 to period T (where T is not very large), we are concerned with the possible points on the y_2-y_2-t diagram. The most rigid category 1 would be a surface on this diagram. This means that, for any given t, the possible combinations of y_{1t} and y_{2t} will lie on a curve on the y_1-y_2 plane. As an example, there may be a structural equation, or an equation resulting from eliminating other endogenous variables from several structural equations, which relates y_{1t}, y_{2t}, and other variables not subject to government control, either directly or indirectly. Again, category 1 is a very special case. By varying the time paths of the policy instruments, one may obtain combinations of y_{1t}, y_{2t}, and t which are not confined to a surface. However, not all points in the y_1, y_2, and t space are reachable by manipulation of government policies. There may be a set of surfaces serving as the lower boundaries for the possible time paths for y_1 and y_2 in the following sense. For any t, one cannot reduce y_{2t} without increasing y_{1t} *or* y_{2s} *or* y_{1s} for $s \neq t$. Thus it may be possible to reduce both y_{1t} and y_{2t}, but some y_{1s} or y_{2s} in another period s will have to increase—otherwise, the surface is not a lowest boundary possible. Category 2, where such surfaces exist, is considered more likely than both category 1 and category 3 where such boundary surfaces do not exist.

To obtain a path for y_1 and y_2 on such boundary surface, we minimize the loss function

$$(2) \quad \sum_{t=1}^{T} [k_{1t}(y_{1t} - a_{1t})^2 + k_{2t}(y_{2t} - a_{2t})^2]$$

subject to the constraint of the dynamic econometric model. Consider the 2T-dimensional space with y_{1t}, y_{2t} (t = 1, 2, ..., T) measured along its coordinates. The points in this space having the same loss are ellipsoids. Contracting one such ellipsoid while keeping the y_1 and y_2 paths attainable by the dynamic econometric model guaran-

tees that the attainable point with minimum loss lies on the boundary surface. To keep the unemployment path close to 5 percent, say, and to find a best inflation path consistent with the econometric model, one may choose $a_{1t} = 5$, $a_{2t} = 0$ (or a small number), $k_{1t} = 1000$, and $k_{2t} = 1$ for t = 1, 2, ..., T. A numerical method for minimizing a quadratic loss function subject to the constraint of a non-linear econometric model can be found in Chow (1975, sec. 12.1), and will not be described here. It will be used for the calculations reported in Sections III and IV below.

Once several optimal paths are obtained, with y_{1t} aimed at 4, 5, 6, and 7 percent, respectively, for instance, one may wish to summarize these paths along the best tradeoff boundary in a two-dimensional diagram. One way to do so is to plot the mean inflation rate $(\Sigma_t y_{2t}/T)$ against the mean unemployment rate $(\Sigma_t y_{1t}/T)$ over the T periods. This would imply a constant rate of substitution between y_{1t} and y_{1s}, and between y_{2t} and y_{2s}. Such an implication would violate the specification of a quadratic loss function which measures the loss by the sums of squared deviations of the economic variables from their targets and not by their sums or their arithmetic means. A second way is to plot $[\Sigma_t(y_{2t} - a_{2t})^2/T]^{1/2}$ against $[\Sigma_t(y_{1t} - a_{1t})^2/T]^{1/2}$ which would penalize the increase of each variable by the square of its deviation from target. Note that when a two-dimensional diagram is used in the multiperiod case, the index number problem cannot be avoided.

For the very long run, if any one ever cares for such an analysis, we can define the equilibrium y_1 and y_2 combination as the constant values towards which these two variables approach as T increases in the multiperiod minimization problem specified above, if such constant values exist. Of course, y_{1t} and y_{2t} might not approach constant values as t increases. Since we treat the very long-run problem in the same way as the above intermediate-run problem by simply increasing the planning horizon T, we can still use such indices as $(\Sigma_t y_{it}/T)$ and $[\Sigma_t(y_{it} - a_i)^2/T]^{1/2}$ for $i = 1, 2$ even if y_{it} itself might not approach a limit as t increases. It is important to note that, as in the case of the intermediate-run problem, optimization is required to obtain the most favorable long-run tradeoff. As long as there exist bad policies which would create more inflation in the long run without improving the unemployment situation, the time paths for y_{1t} and y_{2t} do not all fall on a rigid surface, and one needs to minimize in order to obtain a path on the lower boundary of all feasible paths.

III. Analysis of the St. Louis Model

The St. Louis Model of Andersen and Carlson is well-known. It has an equation explaining money *GNP* by the current and lagged values of money supply *M* and high-employment federal expenditures *E*. Demand pressure and expected price change will help determine the change in the price level. Since both policy instruments *M* and *E* affect the economy through the same money *GNP* variable, in effect the two instruments are only a single policy variable. Mathematically the 2 × 2 matrix of partial derivatives of the unemployment rate y_1 and the rate of price change y_2 with respect to these two instruments has rank 1. If the analysis is limited to one quarter, the St. Louis Model therefore implies a rigid tradeoff relation between y_1 and y_2.

However, in a multiperiod setting, in so far as inflation is affected by expected price change which is determined by past price changes and by the demand pressure which is influenced by the course of real output, alternative paths for the money supply can affect both the expected price change and the demand pressure, thus influencing the paths of inflation and unemployment. There is no reason to expect that arbitrary paths for *M* (or *E*) will yield paths for y_1 and y_2 on the lowest boundary for the unemployment-inflation tradeoff through time. Using the twenty quarters from 1971I–1975IV, and the loss function

$$(3) \quad \sum_{t=1}^{20} k_1(y_{1t} - a_1)^2 + \sum_{t=1}^{20} k_2(y_{2t} - a_2)^2$$

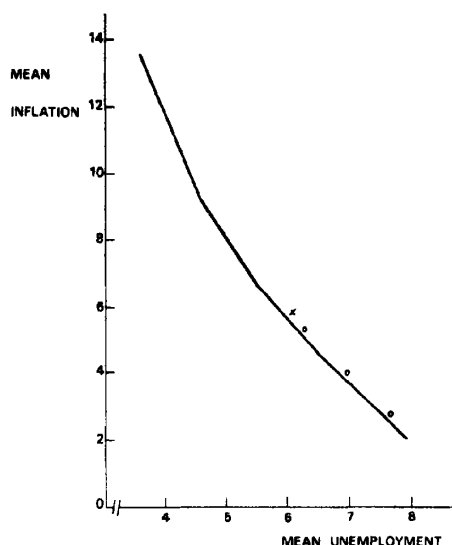



FIGURE 1. TRADEOFF FROM THE ST. LOUIS MODEL

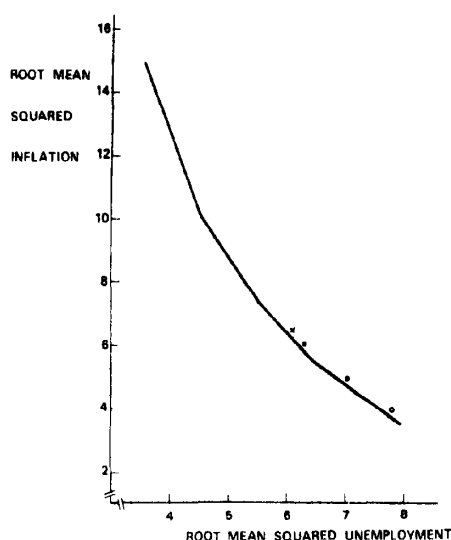



FIGURE 2. TRADEOFF FROM THE ST. LOUIS MODEL

with $k_1 = 10000$, $k_2 = .01$, $a_2 = 2$, and $a_1 =$ 3.5, 4.5, 5.5, 6.5, 7.0, and 8.0, respectively, and letting E_t follow its historical path, we have obtained six optimal paths by minimizing (3) with respect to M_t. $(\Sigma_t y_{2t})/T$ is plotted against $(\Sigma_t y_{1t})/T$ in Figure 1, and $(\Sigma_t y_{2t}^2/T)^{1/2}$ against $(\Sigma_t y_{1t}^2/T)^{1/2}$ in Figure 2. The six optimal points are joined by a solid line. The corresponding points summarizing the paths for y_1 and y_2 resulting from the historical path for M_t are marked by a cross. The points resulting from using a constant percentage growth for M_t of 2, 4, and 6 percent are marked by small circles.

Note that the optimal points dominate the points resulting from the historical path and the smooth growth paths for M. Figure 1 shows that a mean unemployment rate of 4 percent is associated with about 12 percent inflation, and that a 7 percent unemployment corresponds to about 3.6 percent inflation. It would be of interest to examine the dynamic characteristics of the optimizing paths for M, but space limitation prevents an adequate discussion. Suffice it to say that the optimizing paths for M exhibit sizable fluctuations. To inhibit large fluctuations, we can add a term $\Sigma_{t=1}^{20} k_3(M_t - a_{3t})^2$ in the loss function. Minimization of such a function, again using $a_1 =$ 3.5, 4.5, 5.5, 6.5, 7.0, and 8.0, respectively, will yield a curve above the solid curves in Figures 1 and 2. This curve can also be used to define the best tradeoff relationship between y_1 and y_2, under the assumption that fluctuations in the instrument are also penalized.

IV. Analysis of the Michigan Quarterly Econometric Model

Our brief discussion using the Michigan Quarterly Econometric Model by Hymans and Shapiro follows closely the analysis for the St. Louis Model, except that two instruments are used. The instruments are unborrowed reserves UR and nondefense federal expenditures GFO. With more than one instrument, an optimal path along the lowest boundary for the dynamic unemployment-inflation tradeoff is obtained not simply by a suitable dynamic pattern for the one and only control variable (as in the case of the St. Louis Model), but by an optimal combination of the time paths for the instruments. However, variations of the first

instrument UR are inhibited by the inclusion of UR in the loss function, its target values being assumed to follow the historical path. The loss function is

$$(4) \quad \sum_{t=1}^{17} k_1(y_{1t} - a_1)^2 + \sum_{t=1}^{17} k_2(y_{2t} - a_2)^2 + \sum_{t=1}^{17} k_3(UR_t - a_{3t})^2$$

with $k_1 = 10{,}000$, $k_2 = 1.0$, $k_3 = .1$, $a_2 = 2.0$, and $a_1 = 3.5$, 4.5, 5.5, and 6.0, respectively. The period covered is from 1971I 1975I, with T = seventeen quarters. In the calculations, the residuals of the structural equations were given their estimated values so that application of the historical values of the instruments would reproduce the historical paths of the endogenous variables. This is the approach adopted by the *NSF-NBER* Seminar on Comparison of Econometric Models chaired by Lawrence Klein in its optimal control experiments. Figures 3 and 4 have been obtained in the same way as Figures 1 and 2. The four optimal points are joined by a solid line. The corresponding points summarizing the paths

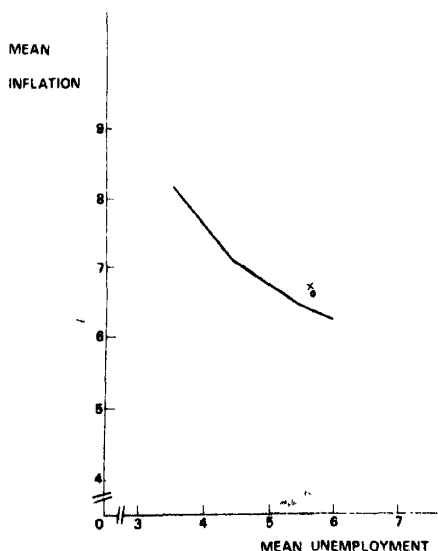



FIGURE 3. TRADEOFF FROM THE MICHIGAN MODEL

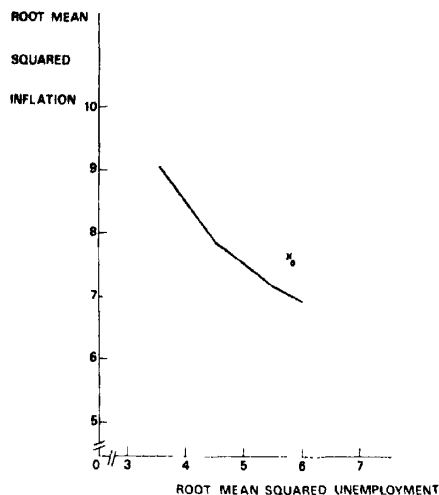



FIGURE 4. TRADEOFF FROM THE MICHIGAN MODEL

for y_1 and y_2 resulting from using the historical values of UR_t and GFO_t are marked by a cross. The points indicated by a circle result from letting UR_t and GFO_t grow at an annual rate of approximately 4.5 and 16.5 percent, respectively. These are the average historical growth rates.

The optimal points clearly dominate the points resulting from the actual as well as the smoothed historical paths for the instruments. The optimal path for *GFO* exhibits some fluctuations. In contrast with the results of Figure 1 for the St. Louis Model, Figure 3 shows that a mean unemployment rate of 4 percent is associated with only 7.6 percent inflation rather than 12 percent. The Michigan tradeoff curve is much flatter than the St. Louis tradeoff curve. The two curves cross at an unemployment rate of about 5.5 percent, with the corresponding inflation rate being 6.5. When the unemployment rate reaches 6.0 percent, the associated inflation rate according to the Michigan Model is 6.3 percent, higher than the 5.6 percent according to the St. Louis Model.

V. Concluding Remarks

We have proposed a definition of the best unemployment-inflation tradeoff and a

method of deriving it numerically from an econometric model. The notion is explained in both a static and a dynamic setting. The St. Louis Model and the Michigan Quarterly Econometric Model have been used to derive the proposed tradeoff relationships. In both cases, it has been shown that the outcomes of inflation and unemployment resulting from other than optimum values of the policy instruments are dominated by the results obtained by optimization. The examples illustrate clearly the need for optimization in order to ascertain the best possible tradeoff. If the optimizing paths of the instruments according to a given econometric model fluctuate too violently for actual implementation, it may be reasonable to define and derive the best tradeoff relationships by penalizing and inhibiting the instability in the instruments. The method proposed could also be employed to compare different econometric models in terms of the most favorable tradeoff relationships which they imply and of the characteristics of the required time paths of the instruments.

At the present time, this use is probably more important than the use of the method proposed in the actual formulation of economic policy because of the possible weaknesses of the current generation of econometric models. For example, the tradeoff relationship shown in Figures 1-4 for the St. Louis and the Michigan models are distinctly different. Policymakers, who have difficulties in reconciling these differences, however, may wish to consult a related work by Chow (1977) on the use of imperfect econometric models for economic policy decisions.

REFERENCES

L. C. Andersen and K. M. Carlson, "A Monetarist Model for Economic Stabilization," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1970, *52*, 7-25.

——, "An Econometric Analysis of the Relation of Monetary Variables to the Behavior of Prices and Unemployment," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972, 166-83.

R. G. Bodkin, "Wage and Price Formation in Selected Canadian Econometric Models," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972, 369-85.

Gregory C. Chow, *Analysis and Control of Dynamic Economic Systems*, New York 1975.

——, "An Approach to the Feedback Control of Nonlinear Econometric Systems," *Annals Econ. Soc. Measure*, Summer 1976, *5*, 297-309.

——, "Usefulness of Imperfect Models for the Formulation of Stabilization Policies," *Annals Econ. Soc. Measure*, Spring 1977, *6*, 175-87.

G. de Menil and J. J. Enzler, "Prices and Wages in the FR-MIT-Penn Econometric Model," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972, 277-308.

Otto Eckstein, *The Econometrics of Price Determination*, Washington 1972.

A. A. Hirsch, "Price Simulations with the OBE Econometric Model," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972, 237-76.

S. H. Hymans, "Prices and Price Behavior in Three U.S. Econometric Models," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972, 309-24.

—— **and H. T. Shapiro,** "The Michigan Quarterly Econometric Model of the *U.S.* Economy," in *The Economic Outlook for 1973*, papers presented to the Twentieth Anniversary Conference, Univ. Michigan 1973.

A. W. Phillips, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957," *Economica*, Nov. 1958, *25*, 283-99.

Dynamic Instability of a Mixed City in the Presence of Neighborhood Externalities

By TAKAHIRO MIYAO*

It has been often asked whether a city, currently accommodating a variety of economic, social, and racial groups, will continue to be a "mixed" city or will become exclusively occupied by a single group, particularly poor blacks, with all other groups moving out of the city in the long run. In order to answer this question properly, it seems essential to develop a long-run dynamic model of a city with emphasis on socioeconomic factors like neighborhood externalities within or among various groups of residents as well as spatial factors like the individual choice of residential space and location. It might be safely said that in the literature the only theoretical work which employs a long-run dynamic approach to this kind of problem is a paper on dynamic models of segregation by Thomas Schelling (1971).[1] Although his socioeconomic analysis has revealed some of the dynamic properties of a racially mixed area with neighborhood externalities, his model itself is quite unsatisfactory from the economic point of view, since it lacks individual utility functions and thereby disregards the aspect of the individual choice of space and location within the city.[2] As a result, he failed to analyze the effect of spatial segregation within the city on the long-run nature of the residential composition of the city.

In this paper I develop a long-run dynamic model of an open city which accommodates two groups of households (for example, blacks and whites, rich and poor, young and old) with their utility functions depending on a consumption good and residential space. I also assume two kinds of neighborhood externalities. The first is called "negative intergroup externalities"; that is, the utility level of a typical household in one group is adversely affected by an increase in the number of households in the other group living in the city. The second is called "positive intragroup externalities"; that is, the utility level of a typical household in one group is positively affected by an increase in the size of its own group in in the city.

First, I define a "mixed-city" equilibrium as a long-run equilibrium which allows the two groups to coexist in the city in the long run, and then introduce a dynamic adjustment process of household movement into and out of the city. It is proved that the mixed-city equilibrium is always unstable in the presence of negative intergroup externalities and/or positive intragroup externalities, if residential land is perfectly homogeneous and there is no differential transport cost incurred within the city. It turns out, however, that in the case of a monocentric city with positive transport cost, the mixed-city equilibrium may be stable or unstable depending on the degree of neighborhood externalities because of the stabilizing effect of spatial segregation between the two groups within the city.

I. Negative Intergroup Externalities

There are assumed to be two groups of households, say, blacks and whites, living together in a city with N_1 households in group 1 and N_2 households in group 2. In

*Assistant professor, department of economics, University of California-Santa Barbara. I wish to thank Richard Arnott, Franklin M. Fisher, John M Marshall, Perry Shapiro, and an anonymous referee for helpful comments and suggestions An earlier version of this paper was presented at the Western Economic Association Conference in Anaheim, California, June 1977

[1]See especially his bounded-neighborhood model, pp. 167 86. Also see Schelling (1972, 1975).

[2]Although spatial aspects are taken account of in some of the recent work related to the present subject, those models are all completely static. For instance, see Susan Rose-Ackerman and John Yinger.

the city the total amount of residential land is given and fixed exogenously. In this section and the following two sections let us suppose that land is perfectly homogeneous in quality, and transportation cost (more rigorously, differential transportation cost) is negligible in the city so that there is no economic reason for a household to prefer one location to another within the city. Here some negative intergroup externalities are introduced, that is, each group does not like the other. More specifically, the utility of a typical household in one group tends to decrease with an increase in the number of households in the other group living in the city. This means that the utility U_i of a typical household in group i depends not only on the amount of a consumption good z_i and the amount of residential land (space) h_i, but also on the size of the other group N_j in the city:

$$(1)\quad U_i = U_i(z_i, h_i, N_j), \qquad i \neq j, \quad (i, j = 1, 2)$$

where negative intergroup externalities are expressed as

$$(2)\quad \partial U_i/\partial N_j < 0, \qquad i \neq j, \qquad (i, j = 1, 2)$$

As a first approximation, it seems reasonable to assume that for given z_i and h_i neighborhood externalities will affect the level of utility without changing the marginal rate of substitution between z_i and h_i so that the utility function can be written in the following separable form:

$$(3)\quad U_i = F_i(z_i, h_i)E_i(N_j), \qquad i \neq j, \quad (i, j = 1, 2)$$

where

$$(4)\quad E_i'(N_j) < 0, \qquad i \neq j, \qquad (i, j = 1, 2)$$

and the function F_i is assumed to possess all the desirable properties like twice differentiability and concavity as usually assumed in the literature.

Given a certain amount of income $w_i > 0$, each household maximizes its utility (3) subject to the budget constraint $z_i + rh_i = w_i$, where r is land rent and the price of the consumption good is given exogenously and normalized as unity.[3] Then the corresponding indirect utility function V_i takes the form

$$(5)\quad V_i = G_i(r, w_i)E_i(N_j), \qquad i \neq j, \quad (i, j = 1, 2)$$

from which it follows that[4]

$$(6)\quad \partial V_i/\partial r = (\partial G_i/\partial r)E_i(N_j) < 0, \quad \partial V_i/\partial w_i = (\partial G_i/\partial w_i)E_i(N_j) > 0$$

and

$$(7)\quad h_i = -(\partial V_i/\partial r)/(\partial V_i/\partial w_i) = -(\partial G_i/\partial r)/(\partial G_i/\partial w_i) \equiv h_i(r, w_i)$$

With the total amount of residential land L given exogenously, we can equate total supply and demand for land as

$$(8)\quad h_1(r, w_1)N_1 + h_2(r, w_2)N_2 = L$$

If we further assume that for any given $w_i > 0$,[5]

$$(9)\quad \partial h_i/\partial r < 0, \qquad h_i(0, w_i) = \infty, \qquad h_i(\infty, w_i) = 0 \qquad (i = 1, 2)$$

then r is uniquely determined by (8) as a function of N_1 and N_2, given w_1, w_2 and L:

$$(10)\quad r = r(N_1, N_2)$$

with

$$(11)\quad \partial r/\partial N_i = -h_i/\{(\partial h_1/\partial r)N_1 + (\partial h_2/\partial r)N_2\} > 0$$

for $N_1 \geq 0$ and $N_2 \geq 0$, but not $N_1 = N_2 = 0$.

At each moment of time, the city is considered to be "closed" in the sense that group sizes N_1 and N_2 are given whereas utility levels u_1 and u_2 are determined from (5) and (10) as

[3]Here we deal with a small city relative to the national economy so that the price of the consumption good is given exogenously in this city.

[4]See the author and Robert Solow for the properties of V_i in the present context

[5]The first condition in (9) means that residential land is a non-Giffen good, and the last two conditions are in fact satisfied in the case of Cobb-Douglas utility functions which will be assumed later.

$$(12)\quad u_i = u_i(N_1, N_2) = G_i\{r(N_1, N_2), w_i\} \cdot E_i(N_j), \quad i \neq j, \quad (i,j = 1,2)$$

In the long run, however, the city is "open" in the sense that group sizes N_1 and N_2 will be completely adjusted through in- and out-migration so as to equate the utility level u_i of the typical household in group i to a certain utility level $\bar{u}_i$ which is attainable elsewhere outside the city in question.[6] More precisely, a long-run equilibrium is characterized by N_1^* and $N_2^* \geq 0$ such that

$$(13)\quad u_i(N_1^*, N_2^*) \leq \bar{u}_i$$

and

$$\{u_i(N_1^*, N_2^*) - \bar{u}_i\}N_i^* = 0 \qquad (i = 1,2)$$

where $\bar{u}_1$ and $\bar{u}_2$ are given exogenously.[7]

In order to examine the stability property of a long-run equilibrium, we introduce a dynamic adjustment process of household movement into and out of the city as follows. Group size N_i is assumed to be increasing or decreasing gradually over time, due to in-migration or out-migration, according as the utility level $u_i(N_1, N_2)$ is higher or lower than $\bar{u}_i$ at each moment of time:

$$(14)\quad DN_i = \begin{cases} f_i[u_i(N_1, N_2) - \bar{u}_i], & \text{when } N_i > 0, \\ f_i[\max\{0, u_i(N_1, N_2) - \bar{u}_i\}], & \text{when } N_i = 0, \quad (i = 1,2) \end{cases}$$

with $f_i'(\ ) > 0$ and $f_i(0) = 0$, where D denotes differentiation with respect to time. It is clear that (13) is equivalent to $DN_i = 0$ $(i = 1, 2)$.

Let us define a mixed-city equilibrium as $N_1^* > 0$ and $N_2^* > 0$ satisfying (13), i.e.,

$$(15)\quad u_i(N_1^*, N_2^*) = \bar{u}_i \qquad (i = 1,2)$$

Assuming the existence of a mixed-city equilibrium, or in other words, assuming that $\bar{u}_1$ and $\bar{u}_2$ are given in such a way that a mixed-city equilibrium exists, it can be shown that the mixed-city equilibrium is locally *unstable* in the presence of negative intergroup externalities. By linearizing (14) in a small neighborhood of the equilibrium with $N_1^* > 0$ and $N_2^* > 0$, we find

$$(16)\quad \begin{bmatrix} D(N_1 - N_1^*) \\ D(N_2 - N_2^*) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} N_1 - N_1^* \\ N_2 - N_2^* \end{bmatrix}$$

where $a_{11} \equiv \partial f_1(0)/\partial N_1$, $a_{12} \equiv \partial f_1(0)/\partial N_2$, $a_{21} \equiv \partial f_2(0)/\partial N_1$ and $a_{22} \equiv \partial f_2(0)/\partial N_2$, all variables being evaluated at the equilibrium. As shown in the Appendix, it follows from (4), (6), and (11) that

$$(17)\quad a_{11} < 0, a_{12} < 0, a_{21} < 0, a_{22} < 0, \quad a_{11}a_{22} - a_{12}a_{21} < 0$$

Thus the equilibrium is locally unstable, because a necessary and sufficient condition for stability is that $a_{11} + a_{22} < 0$ and $a_{11}a_{22} - a_{12}a_{21} > 0$, the latter being violated in our present case.

II. Global Analysis in a Cobb-Douglas Case

In this section we proceed further to examine the global stability property of the dynamic system (14) by assuming a Cobb-Douglas utility function:

$$(18)\quad U_i = (z_i)^{a_i}(h_i)^{b_i}(C_i + N_j)^{-c_i}, \quad i \neq j, \quad (i,j = 1,2)$$

where a_i, b_i, c_i and $C_i (i = 1, 2)$ are all positive constants. Notice that $C_i > 0$ because it is reasonable to have a finite utility level for one group $U_i < \infty$ even when no household in the other group is living in the city.

In this special case, the indirect utility function will be

$$(19)\quad V_i = A_i r^{-b_i}(w_i)^{a_i+b_i}(C_i + N_j)^{-c_i}, \quad i \neq j, \quad (i,j = 1,2)$$

where $A_i \equiv (a_i)^{a_i}(b_i)^{b_i}(a_i + b_i)^{-(a_i+b_i)} > 0 \quad (i = 1,2)$

Then we derive

$$(20)\quad h_i = \beta_i w_i / r \qquad (i = 1,2)$$

where $\beta_i \equiv b_i/(a_i + b_i)$ which is the constant proportion of rent payment to in-

[6] For the concept of closed and open cities, see William Wheaton.

[7] Obviously, (13) implies that $N_i^* = 0$ if $u_i < \bar{u}_i$

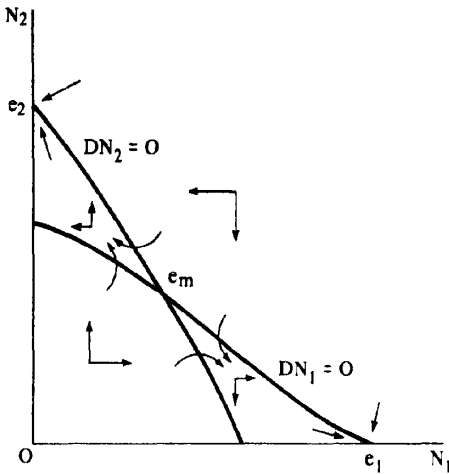



FIGURE 1

come. From (8) together with (20) we can obtain r as

$$r = (\beta_1 w_1 N_1 + \beta_2 w_2 N_2)/L \tag{21}$$

Then, (19) and (21) give

$$u_i(N_1, N_2) = \alpha_i(\beta_1 w_1 N_1 + \beta_2 w_2 N_2)^{-b_i}(C_i + N_j)^{-c_i}, \quad i \neq j, \quad (i,j = 1,2) \tag{22}$$

where $\alpha_i \equiv A_i L^{b_i}(w_i)^{a_i + b_i}$

which is a given constant.

By setting $u_1(\ ) = \bar{u}_1$ and $u_2(\ ) = \bar{u}_2$ alternately, we can find

$$\begin{aligned}(23)\quad 0 < -dN_2/dN_1\,|_{u_1=\bar{u}_1} &= \beta_1 w_1/\{\beta_2 w_2 + c_1(\beta_1 w_1 N_1 + \beta_2 w_2 N_2)/(C_1 + N_2)\} < \beta_1 w_1/(\beta_2 w_2)\\ &< -dN_2/dN_1\,|_{u_2=\bar{u}_2}\\ &= \{\beta_1 w_1 + c_2(\beta_1 w_1 N_1 + \beta_2 w_2 N_2)/(C_2 + N_1)\}/(\beta_2 w_2)\end{aligned}$$

for any N_1, $N_2 > 0$. This means that in Figure 1 the curve representing $DN_2 = 0$ is always steeper than the curve representing $DN_1 = 0$. Then, it follows from the long-run equilibrium condition (13) that whenever there exists a mixed-city equilibrium, say e_m in Figure 1, there must be two other equilibrium points with $N_i^* > 0$ and $N_j^* = 0\ (i, j = 1, 2)$, say e_1 and e_2 in Figure 1, and it also follows from the dynamic adjustment process (14) that as shown in Figure 1 the mixed-city equilibrium e_m is always globally unstable and the city will eventually approach either e_1 or e_2, that is, an "all black" or "all white" city with

$$\begin{aligned}(24)\quad N_i^* &= (\alpha_i)^{1/b_i}(\beta_i)^{-1}(w_i)^{-1}(C_i)^{-c_i/b_i} \cdot (\bar{u}_i)^{-1/b_i},\\ N_j^* &= 0\ (i,j = 1,2)\end{aligned}$$

III. Positive Intragroup Externalities

In this section, there are assumed to be no externalities between the two groups as in the previous sections, but some positive externalities within each group. In other words, the two groups neither like nor dislike each other, but each group likes its own in the sense that the utility of a typical household in one group tends to increase as the size of its own group increases. These kinds of neighborhood externalities may be called positive intragroup externalities and expressed as

$$U_i = F_i(z_i, h_i)E_i(N_i) \qquad (i = 1,2) \tag{25}$$

where

$$E_i'(N_i) > 0 \qquad (i = 1,2) \tag{26}$$

Following Schelling's terminology (1971, p. 165), we might call the utility function (25) "congregationist preferences" meaning that each household wants to congregate with its own group, but is indifferent to the presence of the other group of households in the city. Although congregationist preferences seem less discriminatory than the type of preferences we have assumed in the negative intergroup externality case, it turns out that their long-run effects on the residential composition of the city are equally destabilizing, as suggested by Schelling (1971) in connection with his simple experiments on segregation.

In the present case, the indirect utility function becomes

$$V_i = G_i(r, w_i)E_i(N_i) \qquad (i = 1,2) \tag{27}$$

At each moment of time with N_i given, the

utility level for group i can be written as

$$u_i(N_1, N_2) = G_i\{r(N_1, N_2), w_i\} \cdot E_i(N_i) \qquad (i = 1,2) \tag{28}$$

where $r(\ )$ is defined as (10). Then we can show that the mixed-city equilibrium is locally *unstable* in the presence of positive intragroup externalities. In a small neighborhood of the mixed-city equilibrium, we have the dynamic system which takes the same form as (16). In the present case, the stability condition that $a_{11} + a_{22} < 0$ and $a_{11}a_{22} - a_{12}a_{21} > 0$ cannot be satisfied, because, as shown in the Appendix,

$$a_{11}a_{22} - a_{12}a_{21} < 0 \quad \text{if } a_{11} < 0 \text{ and/or } a_{22} < 0 \tag{29}$$

Note that the condition $a_{11} + a_{22} < 0$ implies $a_{11} < 0$ and/or $a_{22} < 0$.

Finally, it should be added that in the presence of *both* negative intergroup externalities *and* positive intragroup externalities, the instability of a mixed-city equilibrium holds a fortiori, as can be easily shown. A special case of this extended model would be the case with externalities depending on the ratio of group sizes $E_i(N_i/N_j)$ where $E_i'(\ ) > 0$; a form which Schelling (1971) has adopted in his simple analysis of segregation.

IV. The Effect of Segregation in a Monocentric City Case

So far we have assumed that land is perfectly homogeneous in every respect with no differential transport cost incurred within the city. This assumption does not hold, however, in the standard Alonso-Muth-type model of a monocentric city where positive transportation cost is incurred by every resident in commuting from his residence to work in the central business district (*CBD*) (see William Alonso and Richard Muth). In this case, those pieces of land which are closer to the *CBD* have locational advantages in terms of transportation cost over those pieces of land which are located further from the *CBD*. Then the question is how the instability results we have obtained in the previous cases will be modified.

It is well known that in this kind of model with more than one type of household, a "segregated" pattern of residence will emerge in equilibrium: the whole land area will be subdivided into the same number of zones as the number of household types, and each zone is exclusively occupied by a group of households of the same type.[8] If, in particular, there are two groups of households having identical preferences for the consumption good and residential land, but having different income levels, then the lower-income group can be shown to occupy the zone closer to the *CBD* while consuming less land than the higher-income group in equilibrium. Notice that this kind of segregation occurs for economic reasons and is not caused by the presence of neighborhood externalities. However, spatial segregation has a certain effect on the long-run dynamic property of a mixed-city equilibrium in the presence of negative intergroup externalities or positive intragroup externalities. Its effect turns out to be stabilizing, and as a result the dynamic adjustment process may be stable or unstable, depending on the magnitude of the destabilizing effect of neighborhood externalities relative to the stabilizing effect of spatial segregation within the city.

For the purpose of illustration, it is sufficient to consider the case of negative intergroup externalities with the Cobb-Douglas utility function:

$$U_i = (z_i)^a (h_i)^b (C_i + N_j)^{-c_i}, \quad i \neq j, \qquad (i,j = 1,2) \tag{30}$$

where a, b, c_i and $C_i (i = 1, 2)$ are all positive constants. Furthermore, we assume that the transportation cost T incurred by a household living at distance x from the *CBD* can be written as $T(x) = qx$, where q is a given positive constant. Since in equilibrium all households in the same group must attain the same utility level regardless of their lo-

[8] For this well-known result, see Edwin Mills, the author, and Solow.

cation,[9] we find

$$(31)\quad u_i = A\{r_i(x)\}^{-b}(w_i - qx)^{a+b} \cdot (C_i + N_j)^{-c_i}, \qquad i \neq j, \qquad (i,j = 1,2)$$

where the expression on the right-hand side is the indirect utility function with

$$A \equiv a^a b^b (a + b)^{-(a+b)}$$

and $w_i - qx$ in place of A_i and w_i, respectively, in (19), and $r_i(x)$ is the "bid rent" of a household in group i at x corresponding to a utility level u_i which is common for all households in group i. In fact, from (31) the bid rent can be expressed as

$$(32)\quad r_i(x) = A^{1/b}(w_i - qx)^{(a+b)/b} \cdot (C_i + N_j)^{-c_i/b}(u_i)^{-1/b} \qquad (i = 1,2)$$

Without loss of generality, let us assume that $w_1 < w_2$. As pointed out above, all households in group 1 will be accommodated in the inner zone, that is, the land area between the *CBD* and the residential boundary between groups 1 and 2. In view of the fact that the reciprocal of the amount of land per household in group 1 at x, $1/h_1(x)$, is equal to the density of households in group 1 at x, we have

$$(33)\quad \int_0^{x_1} \{m(x)/h_1(x)\}dx = N_1$$

where $m(x)$ is the amount of residential land available at distance x from the *CBD*,[10] and x_1 is the distance from the *CBD* to the boundary between groups 1 and 2. Similarly, all households in group 2 will be located in the outer zone, that is, the area between x_1 and the outer city boundary x_2:

$$(34)\quad \int_{x_1}^{x_2} \{m(x)/h_2(x)\}dx = N_2$$

The residential boundary x_1 is endogenously determined by the condition that the overall rent function be continuous at x_1:

$$(35)\quad r_1(x_1) = r_2(x_1)$$

On the other hand, the outer city boundary x_2 is assumed to be given and fixed exogenously.[11]

In order to determine u_1 and u_2 as functions of N_1 and N_2 in this model, we first solve (35) for x_1, using (31).

$$(36)\quad x_1 = (w_2 S_2 - w_1 S_1)/\{q(S_2 - S_1)\} \equiv x_1(u_1, u_2, N_1, N_2)$$

where

$$S_i = (C_i + N_j)^{-c_i/(a+b)}(u_i)^{-1/(a+b)} (i = 1,2)$$

Since the demand for land per household in group i at x is

$$h_i(x) = \beta(w_i - qx)/r_i(x) = B(w_i - qx)^{-a/b} (C_i + N_j)^{c_i/b}(u_i)^{1/b}$$

with $\beta \equiv b/(a + b)$ and $B \equiv \beta A^{-1/b}$, it follows from (33) and (34) that

(37)

$$M_1(u_1, N_2, x_1) \equiv B^{-1}(C_1 + N_2)^{-c_1/b}(u_1)^{-1/b} \cdot \int_0^{x_1} (w_1 - qx)^{a/b} m(x)\,dx = N_1,$$

$$M_2(u_2, N_1, x_1) \equiv B^{-1}(C_2 + N_1)^{-c_2/b}(u_2)^{-1/b} \cdot \int_{x_1}^{x_2} (w_2 - qx)^{a/b} m(x)\,dx = N_2$$

From (36) and (37) we may determine u_1 and u_2 as functions of N_1 and N_2, and calculate $\partial u_i/\partial N_j$ $(i,j = 1,2)$, as shown in the Appendix;

(38)

$$\partial u_1/\partial N_1 = (Z_2 Y - N_2/b)/(\Delta u_2)$$

$$\partial u_2/\partial N_1 = \{b^2 Z_2 Y - c_2 N_2(1 + g)/(1 + C_2/N_1)\}/(\Delta u_1 b^2)$$

$$\partial u_1/\partial N_2 = \{-b^2 Z_1 Y - c_1 N_1(1 + g)/(1 + C_1/N_2)\}/(\Delta u_2 b^2)$$

$$\partial u_2/\partial N_2 = (-Z_1 Y - N_1/b)/(\Delta u_1)$$

where $Z_1 \equiv \partial M_1/\partial x_1 > 0$, $Z_2 \equiv \partial M_2/\partial x_1 < 0$ (from (37)),

[9] Otherwise, some household could attain a higher utility level by moving to a better location in the city.

[10] Note that $m(x) = 1$ in the linear city case and $m(x) = 2\pi x$ in the circular city case. In what follows we assume that $m(x) > 0$ for all $0 \leq x \leq x_2$.

[11] Alternatively, we may assume that the opportunity cost of land is zero so that we can determine x_2 by $r_2(x_2) = 0$, that is, $x_2 = w_2/q$.

$$Y \equiv (w_2 - w_1)S_1S_2/\{(a + b)q \cdot (S_2 - S_1)^2\} > 0$$

$$g \equiv bY\{(Z_1/N_1) - (Z_2/N_2)\} > 0$$

and $\Delta \equiv N_1N_2(1 + g)/(u_1u_2b^2) > 0$.

Now we can examine the stability property of the dynamic system (16) in the present case with (38), i.e.,

$$(39) \quad \begin{aligned} a_{11} &= f_1'(0)\partial u_1/\partial N_1 < 0 \\ a_{21} &= f_2'(0)\partial u_2/\partial N_1 < 0 \\ a_{12} &= f_1'(0)\partial u_1/\partial N_2 < 0 \\ a_{22} &= f_2'(0)\partial u_2/\partial N_2 < 0 \end{aligned}$$

Furthermore, by rearranging terms, we find

$$(40) \quad a_{11}a_{22} - a_{12}a_{21} = \frac{f_1'(0)f_2'(0)(1 + g)}{\Delta^2 u_1u_2b^2} \Bigg[Z_1YN_2\left(\frac{b}{1+g} - \frac{c_2}{1 + C_2/N_1}\right) - Z_2YN_1\left(\frac{b}{1+g} - \frac{c_1}{1 + C_1/N_2}\right) + \frac{(1+g)(N_2N_1)}{b^2}\left(\frac{b^2}{(1+g)^2} - \frac{c_1}{1 + C_1/N_2}\frac{c_2}{1 + C_2/N_1}\right)\Bigg]$$

It is clear from (39) and (40) that the mixed-city equilibrium is locally *stable* if

$$(41) \quad \frac{c_i}{1 + C_i/N_j} < \frac{b}{1 + g} \qquad (i = 1,2)$$

because in this case we have $a_{11} + a_{22} < 0$ and $a_{11}a_{22} - a_{12}a_{21} > 0$. On the other hand, it is locally *unstable* if

$$(42) \quad \frac{c_i}{1 + C_i/N_j} > \frac{b}{1 + g} \qquad (i = 1,2)$$

since this implies $a_{11}a_{22} - a_{12}a_{21} < 0$.

A possible economic interpretation of these conditions is as follows. Suppose initially u_1 is smaller than $\bar{u}_1$. The resultant out-migration of group 1 will make the city more attractive for the other group, and thus induce more in-migration (or less out-migration) of group 2. This in turn tends to lower u_1 due to negative intergroup externalities, and u_1 will diverge from $\bar{u}_1$ over time. On the other hand, owing to spatial segregation, a decrease in N_1 tends to raise the utility level for group 1 by increasing the amount of land to be occupied by each household in group 1, although this effect is partially offset by the inward movement of the residential boundary between the two groups as a result of a decrease in the land rent to be paid by a smaller number of households in group 1. Thus, the mixed-city equilibrium is stable or unstable, according as the former destabilizing effect $c_i/(1 + C_i/N_j)$, that is, the elasticity of U_i with respect to N_j, is smaller or greater than the latter stabilizing effect as represented by $b/(1 + g)$, that is, the elasticity of U_i with respect to h_i. Note that the presence of the "discount factor" $(1 + g)$ is due to the offsetting effect of the boundary movement. Figures 2 and 3 illustrate stable and unstable cases, respectively. It should be noted that a similar result can be obtained in the case of positive intragroup externalities.

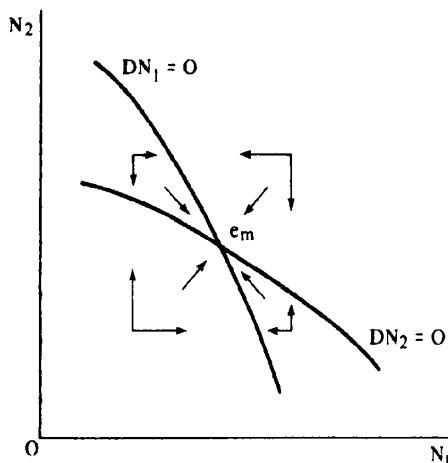



FIGURE 2

V. Concluding Remarks

I have shown that in the presence of negative intergroup externalities or positive intragroup externalities, 1) the mixed-city equilibrium is always unstable if land is

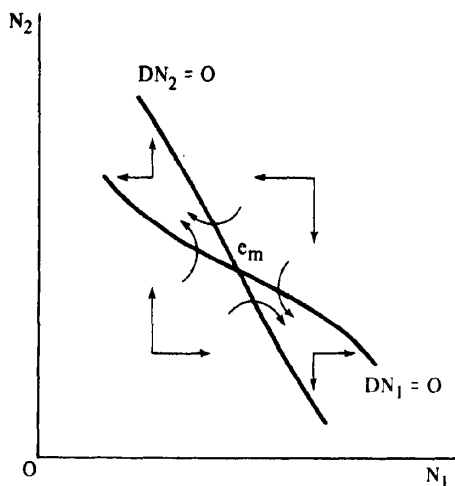



FIGURE 3

perfectly homogeneous in the city with no differential transport cost incurred so that there is no spatial segregation for economic reasons, and 2) the mixed-city equilibrium is not necessarily unstable in the case of a monocentric city with positive transport cost where spatial segregation occurs within the city. In fact, in the latter case the equilibrium is locally stable if the degree of externalities is sufficiently small relative to the elasticity of utility with respect to land.

A few remarks are now in order. First of all, the present analysis can be easily extended to include more complicated cases with any mixture of four kinds of externalities, namely, positive intergroup, negative intergroup, positive intragroup, and negative intragroup. It seems obvious that positive intergroup and negative intragroup externalities have some stabilizing effects, while negative intergroup and positive intragroup externalities are destabilizing as I have analyzed in this paper. If, for instance, group 1 likes group 2 which dislikes group 1, the mixed-city equilibrium is stable or unstable, depending on whether the degree of positive intergroup externalities on the part of group 1 is higher or lower than the degree of negative intergroup externalities on the part of group 2.

Secondly, I have dealt only with a kind of segregation which results from the economic behavior of the household choosing more (less) residential land in exchange for higher (lower) transport cost. However, spatial segregation is expected to be always stabilizing irrespective of its cause or reason. Thus it might be conjectured that the (in)stability results we have obtained in our monocentric city case hold true if spatial segregation within the city results directly from negative intergroup externalities themselves, although a more complicated model would be needed to derive a segregated residential pattern on the basis of negative externalities.

Finally, it should be pointed out that the phenomenon of increasing returns in production can be treated in essentially the same way as positive intragroup externalities, if increasing returns take such a form that the income level (wage rate) for each group is an increasing function of its own group size due to increasing marginal labor productivity. Obviously, in the presence of increasing returns of this kind, the mixed-city equilibrium tends to be unstable: this may appear to be a dynamic version of the well-known proposition of complete specialization in the presence of increasing returns in international trade theory. What is interesting here, though, is that even in the presence of increasing returns, as my analysis suggests, the city may not completely specialize in the long run if there is spatial segregation between two groups of households (workers) in the city and if the degree of increasing returns is sufficiently small relative to the elasticity of utility with respect to residential land.

APPENDIX

Equation (17) follows from (4), (6), and (11), as

$$a_{11} = f_1'(0)(\partial G_1/\partial r)(\partial r/\partial N_1)E_1 < 0$$

$$a_{12} = f_1'(0)\{(\partial G_1/\partial r)(\partial r/\partial N_2)E_1 + G_1E_1'(N_2)\} < 0$$

$$a_{21} = f_2'(0)\{(\partial G_2/\partial r)(\partial r/\partial N_1)E_2 + G_2E_2'(N_1)\} < 0$$

$$a_{22} = f_2'(0)(\partial G_2/\partial r)(\partial r/\partial N_2)E_2 < 0$$

and

$$a_{11}a_{22} - a_{12}a_{21} = -f_1'(0)f_2'(0) \cdot \{G_1E_1'(N_2)(\partial G_2/\partial r)(\partial r/\partial N_1)E_2 + G_2E_2'(N_1)(\partial G_1/\partial r)(\partial r/\partial N_2)E_1 + G_1E_1'(N_2)G_2E_2'(N_1)\} < 0$$

Equation (29) is obtained as

$$a_{11} = f_1'(0)\{(\partial G_1/\partial r)(\partial r/\partial N_1)E_1 + G_1E_1'(N_1)\}$$
$$a_{12} = f_1'(0)(\partial G_1/\partial r)(\partial r/\partial N_2)E_1$$
$$a_{21} = f_2'(0)(\partial G_2/\partial r)(\partial r/\partial N_1)E_2$$
$$a_{22} = f_2'(0)\{(\partial G_2/\partial r)(\partial r/\partial N_2)E_2 + G_2E_2'(N_2)\}$$

and thus

$$a_{11}a_{22} - a_{12}a_{21} = f_1'(0)f_2'(0) \cdot [(\partial G_1/\partial r)(\partial r/\partial N_1)E_1G_2E_2'(N_2) + \{(\partial G_2/\partial r)(\partial r/\partial N_2)E_2 + G_2E_2'(N_2)\} \cdot G_1E_1'(N_1)] = f_1'(0)f_2'(0)[(\partial G_2/\partial r) \cdot (\partial r/\partial N_2)E_2G_1E_1'(N_1) + \{(\partial G_1/\partial r)(\partial r/\partial N_1)E_1 + G_1E_1'(N_1)\} \cdot G_2E_2'(N_2)] < 0$$

if we have $a_{11} < 0$ and/or $a_{22} < 0$.

In order to find (38), first derive from (36) that

$$\partial x_1/\partial u_1 = \{-w_1(S_2 - S_1) + (w_2S_2 - w_1S_1)\} \cdot (\partial S_1/\partial u_1)/\{q(S_2 - S_1)^2\} = -Y/u_1$$
$$\partial x_1/\partial u_2 = \{w_2(S_2 - S_1) - (w_2S_2 - w_1S_1)\} \cdot (\partial S_2/\partial u_2)/\{q(S_2 - S_1)^2\} = Y/u_2$$
$$\partial x_1/\partial N_1 = \{w_2(S_2 - S_1) - (w_2S_2 - w_1S_1)\} \cdot (\partial S_2/\partial N_1)/\{q(S_2 - S_1)^2\} = c_2Y/(C_2 + N_1)$$
$$\partial x_1/\partial N_2 = \{-w_1(S_2 - S_1) + (w_2S_2 - w_1S_1)\} \cdot (\partial S_1/\partial N_2)/\{q(S_2 - S_1)^2\} = -c_1Y/(C_1 + N_2)$$

where $Y \equiv (w_2 - w_1)S_1S_2/\{(a + b)q \cdot (S_2 - S_1)^2\}$. Then, total differentiation of (37) with respect to u_1, u_2, N_1, and N_2 yields

$$\begin{pmatrix} H & J \\ K & L \end{pmatrix}\begin{pmatrix} du_1 \\ du_2 \end{pmatrix} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}\begin{pmatrix} dN_1 \\ dN_2 \end{pmatrix},$$

with

$$H \equiv -N_1/(bu_1) + Z_1\partial x_1/\partial u_1 = -(Z_1Y + N_1/b)/u_1,$$
$$J \equiv Z_1\partial x_1/\partial u_2 = Z_1Y/u_2,$$
$$K \equiv Z_2\partial x_1/\partial u_1 = -Z_2Y/u_1,$$
$$L \equiv -N_2/(bu_2) + Z_2\partial x_1/\partial u_2 = (Z_2Y - N_2/b)/u_2,$$
$$P \equiv 1 - Z_1\partial x_1/\partial N_1 = 1 - c_2Z_1Y/(C_2 + N_1),$$
$$Q \equiv c_1N_1/\{b(C_1 + N_2)\} - Z_1\partial x_1/\partial N_2 = c_1(Z_1Y + N_1/b)/(C_1 + N_2),$$
$$R \equiv c_2N_2/\{b(C_2 + N_1)\} - Z_2\partial x_1/\partial N_1 = -c_2(Z_2Y - N_2/b)/(C_2 + N_1),$$
$$S \equiv 1 - Z_2\partial x_1/\partial N_2 = 1 + c_1Z_2Y/(C_1 + N_2),$$

where $Z_i \equiv \partial M_i/\partial x_1 (i = 1, 2)$. By setting $dN_2 = 0$ and using Cramer's rule, we can solve for $\partial u_1/\partial N_1$ and $\partial u_2/\partial N_1$ as

$$\frac{\partial u_1}{\partial N_1} = \frac{1}{\Delta}\begin{vmatrix} P & J \\ R & L \end{vmatrix} = \frac{1}{\Delta}\left\{\left(1 - \frac{c_2Z_1Y}{C_2 + N_1}\right)\frac{Z_2Y - N_2/b}{u_2} + \frac{Z_1Y}{u_2}\,\frac{c_2(Z_2Y - N_2/b)}{C_2 + N_1}\right\} = (Z_2Y - N_2/b)/(\Delta u_2),$$

$$\frac{\partial u_2}{\partial N_1} = \frac{1}{\Delta}\begin{vmatrix} H & P \\ K & R \end{vmatrix} = \frac{1}{\Delta}\left\{\frac{Z_1Y + N_1/b}{u_1}\,\frac{c_2(Z_2Y - N_2/b)}{C_2 + N_1} + \left(1 - \frac{c_2Z_1Y}{C_2 + N_1}\right)\frac{Z_2Y}{u_1}\right\}$$
$$= \frac{1}{\Delta}\left\{\frac{Z_2Y}{u_1} - \frac{c_2(N_1N_2 + bZ_1N_2Y - bZ_2N_1Y)}{u_1b^2(C_2 + N_1)}\right\} = \frac{b^2Z_2Y - c_2N_2(1 + g)/(1 + C_2/N_1)}{\Delta u_1b^2},$$

where $\Delta \equiv HL - JK = N_1N_2(1 + g)/(u_1u_2b^2)$ and $g \equiv bY\{(Z_1/N_1) - (Z_2/N_2)\}$. Similarly, by setting $dN_1 = 0$ we obtain

$$\frac{\partial u_1}{\partial N_2} = \frac{1}{\Delta}\begin{vmatrix} Q & J \\ S & L \end{vmatrix}$$

$$= \frac{-b^2Z_1Y - c_1N_1(1 + g)/(1 + C_1/N_2)}{\Delta u_2 b^2},$$

$$\frac{\partial u_2}{\partial N_2} = \frac{1}{\Delta}\begin{vmatrix} H & Q \\ K & S \end{vmatrix} = -\frac{Z_1Y + N_1/b}{\Delta u_1}$$

REFERENCES

William Alonso, *Location and Land Use*, Cambridge, Mass. 1964.

Edwin S. Mills, *Urban Economics*, Glenview 1972.

T. Miyao, "Dynamics and Comparative Statics in the Theory of Residential Location," *J. Econ. Theory*, Aug. 1975, *11*, 133–46.

Richard F. Muth, *Cities and Housing*, Chicago 1969.

S. Rose-Ackerman, "Racism and Urban Structure," *J. Urban Econ.*, Jan. 1975, *2*, 85–103.

T. C. Schelling, "Dynamic Models of Segregation," *J. Math. Soc.*, July 1971, *1*, 143–86.

———, "A Process of Residential Segregation: Neighborhood Tipping," in Anthony H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972.

———, "Segregation on a Continuous Variable," in J. Bergsman and H. L. Wiener, ed., *Urban Problems and Public Policy Choices*, New York 1975.

R. M. Solow, "On Equilibrium Models of Urban Location," in Michael Parkin, ed., *Essays in Modern Economics*, London 1973.

W. C. Wheaton, "A Comparative Static Analysis of Urban Spatial Structure," *J. Econ. Theory*, Oct. 1974, *9*, 223–37.

J. Yinger, "Racial Prejudice and Racial Residential Segregation in an Urban Model," *J. Urban Econ.*, Oct. 1976, *3*, 383–96.

The Rest of the World's Offer Curve: Note

By WOLFGANG MAYER*

The diagrammatic tool of the rest of the world (*RW*) offer curve had been employed long before Gary Becker rigorously demonstrated its relationship to individual-country offer curves. Becker's contribution was to remove the then prevailing suspicion[1] that offer curve analysis was limited to a two-country framework and that *RW* offer curves were not legitimate tools of trade theory.

For the generally used variable cost model, Becker draws the basic conclusion that a "...complete justification for the shape of the rest of the world offer curve as usually drawn is provided by assuming that there is an infinite number of countries *and* that there is continuous variation in the no trade lines ... of these countries" (p. 568). Although the expression "as usually drawn" is only defined by a general reference to Francis Y. Edgeworth and Alfred Marshall, Becker (p. 568) leaves little doubt that continuous differentiability of the trade vector's components with respect to the terms of trade, or what he calls smoothness, is the crucial property he has in mind. In fact, the diagram on which Becker's conclusion is based, his Figure 7 (p. 567), reveals clearly that the *RW* offer curve, denoted by *OEF*, is not continuously differentiable at point *E*, given there are only two countries and their no-trade price ratios are distinctly different. At point *E* the aggregate offer curve has a kinked indentation.

The purpose of my note is to point out that, in the variable cost case, neither the "infinite number of countries" nor the "continuous variation in no-trade lines" condition is necessary for continuous differentiability of *RW* offer curves. As long as individual countries' offer curves are differentiable, the *RW* offer curve will be differentiable as well. On the other hand, a second "as usually drawn" property,[2] namely strict concavity of the function relating exports to imports, does not necessarily hold for the rest of the world even though it may hold for each individual country.

Becker's construction of a *RW* offer curve, presented in Sections II (constant cost) and III (variable cost) of his paper, is an application of an aggregation technique he develops in Section I. It should be pointed out first that Becker's aggregate offer curves of Section I are quite different in nature from those of Sections II and III. In Section I he aggregates offer curves of those countries which at given terms of trade export the same commodity. Since the number of countries exporting a given commodity changes with varying terms of trade, the number of countries is not held constant along this kind of aggregate offer curve. Alternatively, one can interpret the curve as portraying *gross* amounts of a commodity offered for export by all countries combined at different terms of trade. Becker's diagrams show quite correctly that, at terms of trade where an additional country becomes an exporter of a given commodity, the aggregate curve has a kinked indentation.[3]

When Becker applies his technique of Section I to the derivation of *RW* offer curves in Sections II and III, he quite correctly states that certain modifications of the technique are in order. Whereas the

*Associate professor of economics, University of Cincinnati.

[1] For an expression of this suspicion, see Lloyd Metzler.

[2] As usually drawn it certainly excludes such phenomena as bow-tie and backward-bending offer curves as found, for example, in Harry Johnson and in Murray Kemp and Ronald Jones.

[3] This aggregate offer curve concept has been applied to customs union theory by Sven Arndt and Sukesh Ghosh. At the end of my note it is briefly demonstrated that Ghosh's attempt to replace these indentations with flat line segments is not correct.

FIGURE 1

number of countries varies along the offer curves of Section I, it is fixed for *RW* offer curves. Also, the latter now shows *net* amounts offered (p. 565). And while the offer curve of Section I is limited to the first quadrant, it runs through the first and third quadrants in the case of a *RW* offer curve (p. 566). Unfortunately, this awareness is not reflected in either the conclusions or the diagrammatic representation (Figure 7, p. 567) of the all-important variable cost case. Both imply that the *RW* offer curve has kinked indentations unless the number of aggregated countries goes to infinity and their no-trade price ratios are continuously variable.

Figure 1 of my note shows the technique which should be employed in deriving the *RW* offer curve as faced by country *A*. Using the same notation as Becker in his Figure 7, I assume that there are two countries *B* and *C*, whose offer curves are described by *BOB* and *COC*, respectively. By drawing the curves through both the first and third quadrants, the possibility of reversals in a country's trade pattern is emphasized. Assuming the world consists of countries *A*, *B*, and *C*, the *RW* offer curve faced by country *A* is obtained by summation of the export-import combinations of *B* and *C* at all possible price ratios. To illustrate this summation procedure for a given price ratio, I take $1/p_1$ and indicate the corresponding export-import combinations of *B* and *C* by points B^1 and C^1, respectively. Their combined export-import combination at $1/p_1$ is indicated by point E^1 which is obtained by subtracting the distance OC^1 from B^1O, using B^1 as the starting point. Similarly, if the price ratio is $1/p_0$, country *C*'s no-trade price ratio, point E^0 will be a point on the combined (as well as country *B*'s) offer curve. Applying the same procedure for all other values of $1/p$, the *RW* offer curve is traced out, as denoted by *ROR*. Just as *BOB* and *COC*, the *ROR* curve is continuously differentiable[4] independent of the number of countries and their no-trade ratios. As to the positioning of *ROR*, it lies to the left and below *BOB* for all $1/p > 1/p_0$, above the *COC* line for all $1/p < 1/p_2$, and between *BOB* and *COC* for all $1/p_0 > 1/p > 1/p_2$.

Another as usually drawn property of the offer curve concerns its curvature. Generally, the curve is drawn strictly concave from below as exemplified by the two individual countries' offer curves, *BOB* and *COC* in Figure 2. Aggregation of *BOB* and *COC* yields a *RW* offer curve *ROR*, which is not necessarily strictly concave from below everywhere. In the diagram, this claim is demonstrated by using the adding-up procedure and notation of Figure 1 for three different terms of trade in the relevant price range. As can be shown, a strictly concave-from-above segment is likely to occur if both curves *BOB* and *COC* have almost no curvature in the relevant price range, and country *B*'s imports expand at a much faster rate than country *C*'s. The conclusion that the basic curvature of the *RW* offer curve is not necessarily the same as

[4]This property follows from the theorem that the sum of two continuously differentiable functions is continuously differentiable.

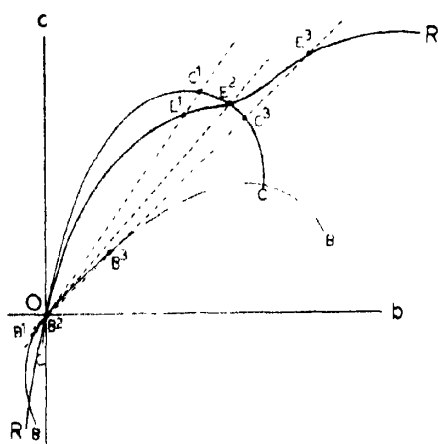
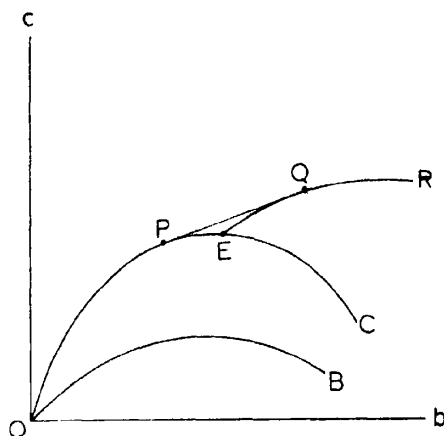



FIGURE 2

FIGURE 3

that of the individual countries' curves is of importance for optimal trade policy. In selecting country A's optimal tariff, care must be taken that the second-order conditions for maximizing utility are satisfied and that a globally, and not just locally, optimal solution is obtained.

So far I have concerned myself only with the variable cost case. It ought to be mentioned here that the proposition that the RW offer curve is continuously differentiable does not carry over to the constant cost model. In the latter case, individual countries' offer curves have flat line segments, implying that the trade vector's components are not continuously differentiable with respect to the terms of trade since no unique limits of the first derivative exist. As one can easily visualize by looking at Figure 2, the RW offer curve would have a kinked indentation at point E^2 if OB had a linear segment starting at the origin. Hence, the drawing of Figure 6 in Becker's paper (p. 565) is correct.

Finally, I return to the aggregation procedure developed by Becker in Section I of his paper. I want to comment briefly on Ghosh's attempt to replace the kinked segment of the resulting aggregate offer curve by a flat line segment. Figure 3 shows such an aggregate offer curve for the variable cost case, with the kink occurring at point E.[5] Ghosh argues that the kinked segment PEQ should be replaced by the straight line PQ, which is obtained by drawing a common tangent to the curves ER and OC. Ghosh argues that efficiency of trading processes requires this modification of Becker's analysis, whereby he employs a technique originally developed by Nicholas Georgescu-Roegen and John Hicks. The problem with Ghosh's approach is that nothing along the straight line PQ represents optimal trading for the two countries involved. The fact that PQ lies above PEQ implies that aggregate exports, which are offered in return for a given amount of imports, are always larger along PQ than along PEQ. If one drew the relevant trade indifference curves of countries B and C into the diagram, one could easily see that in moving from PQ to PEQ at least one country can be made better off without the other country becoming worse off. Georgescu-Roegen's approach is not applicable since in his analysis a higher level of utility can be reached by maximizing output of one commodity for given amounts of the other commodity, whereas utility is reduced when Ghosh maximizes exports for a given level of imports. Ghosh's path is not opti-

[5]The diagram reproduces the essence of Ghosh's Figure 4 (p. 99), but maintains the previously used notation of Becker.

mal in the sense of utility maximization for the countries involved, although it is mathematically optimal in the sense of showing maximum exports for a given level of imports of the two countries.

REFERENCES

S. W. Arndt, "Customs Union and the Theory of Tariffs," *Amer. Econ. Rev.*, Mar. 1969, *59*, 108–17.

G. S. Becker, "A Note on Multi-Country Trade," *Amer. Econ. Rev.*, Sept. 1952, *42*, 558–68.

Francis Y. Edgeworth, *Papers Relating to Political Economy*, Vol. II, London 1925.

N. Georgescu-Roegen, "Leontief's System in the Light of Recent Results," *Rev. Econ. Statist.*, Aug. 1950, *32*, 214–22.

S. K. Ghosh, "Toward a Theory of Multiple Customs Unions," *Amer. Econ. Rev.*, Mar. 1974, *64*, 91–101.

J. R. Hicks, *Value and Capital*, Oxford 1965.

H. G. Johnson, "International Trade, Income Distribution, and the Offer Curve," *Manchester Sch. Econ. Soc. Stud.*, Sept. 1959, *27*, 241–60.

M. C. Kemp and R. W. Jones, "Variable Labor Supply and the Theory of International Trade," *J. Polit. Econ.*, Feb. 1962, *70*, 30–36.

Alfred Marshall, *Money, Credit and Commerce*, London 1923, Appendix J.

L. A. Metzler, "Graham's Theory of International Values," *Amer. Econ. Rev.*, June 1950, *40*, 301–22.

Factor-Price Uncertainty with Variable Proportions

By MARION B. STEWART*

Unless managers are risk indifferent, an uncertain operating environment is likely to have important effects on a firm's output and factor-proportion decisions. Recent papers by Agnar Sandmo, Hayne Leland, and Duncan Holthausen, among others, have examined the various effects of demand uncertainty; but these studies have not considered the effects of *input* price uncertainty. As shown below, factor-price uncertainty may have important effects on firms' factor-proportion decisions, leading risk-averse managers to substitute fixed (riskless) factors of production for factors such as raw materials whose prices are subject to random fluctuation

Consider a hypothetical firm which uses two inputs, X_1 and X_2, to produce a "final product." We shall suppose that the firm purchases input X_1 (raw materials) in a competitive market at price r_1 per unit and combines that input with X_2 (plant and equipment) to produce its final product. The profit in any period t is $\pi = pq - r_1x_1 - r_2x_2$, where p is output price, q is quantity sold (= quantity produced), r_2 is the unit cost (including opportunity costs) of input X_2, and x_1 and x_2 are the quantities of X_1 and X_2 required to produce q units of output.

Let us suppose that input substitution is possible, but we also assume that the plant configuration is not instantaneously adjustable, so that x_2 must be chosen at some time prior to the beginning of period t. An important assumption is that the price of raw materials is subject to random fluctuation, so that r_1 is not known at the time x_2 is chosen. Suppose, however, that r_1 has a subjective probability density $\phi(r_1)$, known to the firm's manager at the time x_2 must be determined

It is assumed initially that p, q, and r_2 are known with certainty, and that the firm's manager's objective is to choose the plant configuration x_2 which maximizes $EU(\pi)$ where U is a von Neumann-Morgenstern utility function with profit as its argument, and E is the expected value operator.

I. Optimal Input Choices under Factor-Price Uncertainty

Assume that the production function $q = f(x_1, x_2)$ can be solved for the raw materials requirement

$$x_1 = g(x_2, q) \tag{1}$$

where $g(x_2, q)$ is the quantity of X_1 necessary to produce output q with plant configuration x_2. The assumption that X_1 and X_2 can be substituted for one another to produce a given quantity of output implies that $\delta g/\delta x_2 < 0$; if the isoquants are convex, we also have $\delta^2 g/\delta x_2^2 > 0$. Hence $\pi = pq - r_1 g(x_2, q) - r_2 x_2$, and a necessary condition for utility maximization is[1]

$$E[U'(\pi)(-r_1\delta g/\delta x_2 - r_2)] = 0 \tag{2}$$

If the firm's manager is risk neutral, then $U'(\pi)$ is a constant number and (2) may be written as $U'(\pi)(-Er_1\delta g/\delta x_2 - r_2) = 0$ or

$$-\delta g/\delta x_2 = r_2/Er_1 \tag{3}$$

The risk-neutral firm equates the marginal rate of technical substitution to the expected factor-price ratio, thus minimizing expected cost.

A manager with a non-linear utility function, however, will depart from this expected cost-minimization strategy, as it is now shown. Define $A \equiv U'(\pi) > 0$, and $B \equiv -r_1\delta g/\delta x_2 - r_2$, so that we may write (2) as

*Assistant professor, Rutgers College

[1]The assumption of convex isoquants insures that equation (2) is decreasing in x_2, and hence satisfies the second-order conditions for a maximum, provided U is linear or concave

$E(AB) = EA\,EB + cov(A,B) = 0$. If the firm's manager is risk averse, then U is concave and A is an increasing function of r_1. Since B is clearly an increasing function of r_1, we then have $cov(A,B) > 0$, so that (2) is satisfied only if $EB < 0$. Hence a risk-averse manager sets $-\delta g/\delta x_2 < r_2/Er_1$, using more plant and less raw material than does a risk-neutral manager.[2]

This result appears to be quite general, and holds, for example, in the case in which more than one input is subject to factor-price uncertainty.[3] I show in the Appendix, Section A that in the multiple input case, a risk-averse firm uses less of *every* risky input and more of the riskless input(s) than does a risk-neutral firm.

Since raw material prices are subject to random fluctuation, a risk-averse manager protects himself against the risk of very low profits by overinvesting in fixed-cost plant and equipment. If raw material prices are lower than expected, the firm will be too "capital intensive" to benefit fully from the low material costs; but if raw material prices are higher than expected, the manager is protected against losses or very low profits. As expected, the risk-averse manager foregoes the opportunity of earning large profits in order to insure against large losses. A risk-preferring manager, on the other hand, underinvests in plant and increases the likelihood of both very high profits and large losses.

II. Increasing Aversion to Risk

I have argued that a risk-averse manager faced with factor-price uncertainty will use less of a risky input than would a risk-neutral manager, thus departing from the input-use ratio which minimizes expected cost. A more general statement of the relationship between risk aversion and factor proportions is possible, as well: it is argued below that as aversion to risk increases, a manager will use less of a risky input and correspondingly more of a riskless input.

The Pratt-Arrow coefficient of (absolute) risk aversion is defined by $\alpha(\pi) \equiv -U''(\pi)/U'(\pi)$. The coefficient α is positive for a risk-averse firm, and the degree of risk aversion—as measured by the amount a firm would voluntarily pay to insure itself against a risk of given magnitude—increases as α increases. Using this notation, it is easily shown that as α increases, the optimal input-use ratio x_2/x_1 increases. Since a risk-neutral firm chooses the ratio which minimizes expected cost, increasing risk aversion leads to increasing overinvestment in plant and increasing underutilization of raw materials.[4] A sketch of the proof is given in the Appendix, Section B.

It is widely believed that rational firms and households should exhibit decreasing absolute risk aversion, that α should be a decreasing function of wealth. If firms do in fact exhibit decreasing absolute risk aversion, then it follows that a risk-averse firm will lower its optimal input-use ratio x_2/x_1 as its wealth increases. Somewhat informally, this suggests that larger (risk-averse) firms will operate closer to the cost-minimizing input-use ratio than will smaller firms, *ceteris paribus*.

III. The Effect of Increased Risk on Optimal Input Choices

We may examine the effect of increased uncertainty about input prices by defining the material price r_1 as

$$(4) \qquad r_1 \equiv \gamma v + \theta, \qquad v > 0$$

where v is a random variable, and γ and θ are nonstochastic "shift parameters." We

[2]That is, a risk-averse firm will produce a given quantity of output with a higher input-use ratio x_2/x_1 than would be chosen by a risk-neutral firm. If the output quantity q can be varied, a risk-averse firm will not in general wish to produce the same output quantity as a risk-neutral firm, but if the production function $f(x_1, x_2)$ is homogeneous, then a risk-averse firm will operate with a higher input-use ratio than will a risk-neutral firm, regardless of the output quantity produced (since in that case the risk-neutral firm's optimal input-use ratio will be independent of q).

[3]I am grateful to Michael Taussig for raising this point.

[4]In the general case, increasing risk aversion leads to decreased use of *all* risky inputs and increased use of riskless inputs. See the Appendix, Section B.

may then increase the uncertainty about r_1 simply by increasing γ, thus amplifying the variability of ν. So that the expected input price does not change and we observe only the effect of increased risk,[5] we may insure that $d(Er_1) = dE(\gamma\nu + \theta) = E\nu d\gamma + d\theta = 0$ by choosing θ such that $d\theta/d\gamma = -E\nu$.

Now substitute (4) into the first-order condition (2) and differentiate (2) with respect to γ to obtain

$$(5)\quad E(-U''(\pi)x_1(\nu - E\nu)(-r_1\delta g/\delta x_2 - r_2)) + E(-\delta g/\delta x_2 U'(\pi)(\nu - E\nu))$$

It is shown in the Appendix, Section C, that under the hypothesis of decreasing absolute risk aversion, (5) is unambiguously positive. Hence a marginal increase in factor-price uncertainty increases (2), requiring an increase in x_2 (and hence a decrease in x_1) to equate (2) to zero. Perhaps not surprisingly, increased factor-price uncertainty further reduces a risk-averse firm's use of the risky input and moves the firm farther from the cost-minimizing input-use ratio. As usual, this result generalizes in the obvious way to the multiple input case.

Futures markets, long-term contracts, and "vertical integration" all serve to reduce uncertainty about factor prices; and we should expect risk-averse managers to take steps to reduce factor-price uncertainty, since the resultant reduction in profit fluctuations will directly increase the expected utility from profits (i.e., $d(EU(\pi))/d\gamma < 0$). For many products, long-term contracts may be sufficient to greatly reduce factor-price uncertainty; but in other cases particularly when factor "quality" is difficult to measure—vertical integration may be more appropriate. By reducing factor-price uncertainty,[6] vertical integration increases a risk-averse manager's utility; and the reduced uncertainty, as we have seen, will also lead on the average to lower production costs. Since efficient resource use is socially desirable, the encouragement of risk-reducing vertical integration may well be sound public policy.

IV. Input Choices and Output Uncertainty

It was assumed above that while factor prices were subject to random fluctuation, both output price p and quantity q were nonstochastic. I now show that output-price uncertainty has no effect on my conclusions, while uncertainty about q has an effect opposite that of uncertainty about r_1.

Suppose that price p is not precisely known at the time x_2 is chosen. In that case, using subscripts to indicate the variable to which the expected value operator applies, we may write the first-order condition (2) as

$$E(U'(\pi)(-r_1\delta g/\delta x_2 - r_2)) = E_{r_1}[E_p[U'(\pi)(-r_1\delta g/\delta x_2 - r_2)| r_1]] = E_{r_1}[(-r_1\delta g/\delta x_2 - r_2)E_p(U'(\pi) \mid r_1)] = 0$$

But the conditional expectation $E_p(U'(\pi) \mid r_1)$ is itself a random variable, increasing in r_1; so my conclusions are completely unaffected by the existence of output-price uncertainty.

If on the other hand the firm fixes its output price prior to period 1 but faces demand uncertainty,[7] the first-order condition is

$$E_{r_1}[E_q[U'(\pi)(-r_1\delta g/\delta x_2 - r_2) \mid r_1]] = 0$$

and further simplification is not possible since $\delta g/\delta x_2$ is a function of q. If X_2 is a noninferior factor, then $(-\delta g/\delta x_2)$ is an in-

[5] This mechanism for effecting a "mean-preserving spread" in a probability distribution is due to Sandmo.

[6] There are two risk-reducing characteristics of vertical integration. First, vertical integration is likely to improve a firm's information about conditions in the upstream market; but uncertainty about the true opportunity cost of the upstream product will be reduced by a vertical merger even if the integrated firm has no better information than was available to the downstream firm before the merger. This somewhat surprising result is due to the observation that the opportunity cost of a manufactured upstream factor is the marginal revenue associated with the price in the upstream market, not the market price itself. Since marginal revenue is less than price in all but perfectly competitive markets, the uncertainty about MR will generally be less than the uncertainty about the market price. On this point, see my paper on transfer pricing under uncertainty.

[7] This implies a less-than-perfectly competitive market structure. By assumption, a competitive firm can sell all it can produce.

creasing function of q; and since $U'(\pi)$ is a decreasing function of q, the conditional expectation $E_q(AB)$ must be less than E_qA E_qB. It follows that a risk-averse manager facing demand uncertainty (with known input prices) will *underinvest* in the fixed-cost input.[8] Since I have argued that factor-price uncertainty will lead to overinvestment in a fixed-cost input, *ceteris paribus*, a risk-averse manager facing both demand uncertainty and factor-price uncertainty may in general either use too much or too little of the fixed-cost input.

V. Summary and Concluding Remarks

I have considered a simple problem in which a hypothetical firm combines a "fixed" input, available at a known factor price, with a "variable" input whose price is subject to random fluctuation, to produce a given output quantity. The major findings are readily summarized:

Compared with a risk-neutral firm, a risk-averse firm uses less of the risky input and more of the fixed-price input, thus deviating from the input-use ratio which minimizes expected cost.

Increasing aversion to risk leads to increasing use of the fixed-price input and decreasing use of the risky input; the more risk averse is a firm's manager, the farther from the least-cost input-use ratio it operates.

A reduction in factor-price uncertainty, due to long-term contracting or vertical integration, increases a risk-averse firm's use of the risky input, hence reducing expected production costs.

These results generalize in an obvious way to the multiple input case: risk-averse firms use less of every risky input (hence more of the fixed-price inputs) than do risk-neutral firms, and further reduce their use of any risky input if uncertainty about that factor price increases.

These conclusions are unaffected by the addition of output-price uncertainty. As noted in Section IV, however, *demand* uncertainty and input-price uncertainty have opposite effects on risk-averse managers' optimal input-use ratios; hence the effect of decreased factor-price uncertainty on the efficiency of price-setting firms which face uncertain demand can not be predicted a priori.

[8]See Holthausen for a detailed discussion of this point.

Appendix

A

The multiple input case is easily evaluated by directly considering the firm's constrained optimization problem. Define the production function $q = f(x_1, x_2, \ldots x_i, \ldots x_n)$, with i risky inputs and $n - i$ fixed (hence riskless) inputs; and form the Lagrangian

$$L = EU(\pi) + \lambda[q - f(x_1, \ldots, x_n)]$$

Optimal use of inputs i and n, for example, requires

$$\delta L/\delta x_i = E[U'(\pi)r_i] - \lambda\delta f/\delta x_i = 0$$
$$\delta L/\delta x_n = r_n EU'(\pi) - \lambda\delta f/\delta x_n = 0$$

or

$$\text{(A1)} \qquad \frac{\delta f/\delta x_i}{E[U'(\pi)r_i]} = \frac{\delta f/\delta x_n}{r_n EU'(\pi)}$$

If U is concave, then $U'(\pi)$ is an increasing function of r_i, and $E[U'(\pi)r_i] > Er_i EU'(\pi)$. Hence the risk-averse manager chooses inputs i and n such that

$$\text{(A2)} \qquad \frac{\delta f/\delta x_i}{Er_i} > \frac{\delta f/\delta x_n}{r_n}$$

Given diminishing returns to each input, (A2) implies that a risk-averse manager uses less of each risky input (and thus more of the riskless inputs) than does a risk-neutral manager.

B

Consider two firms with concave utility functions $U_1(\pi)$ and $U_2(\pi)$, and risk aver-

sion coefficients $\alpha_1(\pi)$ and $\alpha_2(\pi)$; and let firm one be the more risk averse, so that $\alpha_1(\pi) > \alpha_2(\pi) > 0$ for any π. Define r_1^0 as the value of r_1 at which $(-r_1\delta g/\delta x_2 - r_2) = 0$, and let π^0 be the value of π when $r_1 = r_1^0$. Also let x_2^* be the value of x_2 which satisfies the first-order condition (2) for firm one, and let $B(x_2^*) \equiv -r_1\delta g/\delta x_2 - r_2$, evaluated at $x_2 = x_2^*$ (see the discussion in Section I). Partitioning (2) into two parts and dividing by the constant number $U_1'(\pi^0)$, we may then write firm one's first-order condition as

$$\text{(A3)} \quad \int_0^{r_1^0} \frac{U_1'(\pi)}{U_1'(\pi^0)} B(x_2^*)\phi(r_1)dr_1 + \int_{r_1^0}^{r} \frac{U_1'(\pi)}{U_1'(\pi^0)} B(x_2^*)\phi(r_1)dr_1 = 0$$

Now evaluate firm two's first-order condition at x_2^*, and divide by the constant number $U_2'(\pi^0)$ to obtain

$$\text{(A4)} \quad \int_0^{r_1^0} \frac{U_2'(\pi)}{U_2'(\pi^0)} B(x_2^*)\phi(r_1)dr_1 + \int_{r_1^0}^{r} \frac{U_2'(\pi)}{U_2'(\pi^0)} B(x_2^*)\phi(r_1)dr_1$$

Subtracting (A3) from (A4), we have

(A5)

$$\int_0^{r_1^0} \left[\frac{U_2'(\pi)}{U_2'(\pi^0)} - \frac{U_1'(\pi)}{U_1'(\pi^0)}\right] B(x_2^*)\phi(r_1)dr_1 + \int_{r_1^0}^{\infty} \left[\frac{U_2'(\pi)}{U_2'(\pi^0)} - \frac{U_1'(\pi)}{U_1'(\pi^0)}\right] B(x_2^*)\phi(r_1)dr_1$$

Relying on a theorem by John Pratt, it is easy to show that if $\alpha_1(\pi) > \alpha_2(\pi)$, then the first bracketed expression is positive and the second bracketed expression is negative. Since $B(x_2^*) < 0$ for $r_1 < r_1^0$, and $B(x_2^*) > 0$ for $r_1 > r_1^0$, (A5) is clearly negative. Hence (A4) must be negative when $x_2 = x_2^*$; and since (A4) is a decreasing function of x_2, firm two (the less risk-averse firm) must use *less* of the fixed-cost input in order to satisfy its first-order condition for utility maximization.

The same method of analysis can be applied to equation (A1) to show that the more risk-averse firm uses less of each risky input, hence more of the riskless inputs.

C

Since $\delta g/\delta x_2$ is negative, and $U'(\pi)$ is an increasing function of r_1 (and hence of v), the second term in (5) is clearly positive. To determine the sign of the first term, define $C \equiv U'(\pi)(-r_1\delta g/\delta x_2 - r_2)$ so that, using the definition of α, we may write the first term in (5) as

$$\begin{aligned}\text{(A6)}\ x_1E[\alpha U'(\pi)(v - Ev)(-r_1\delta g/\delta x_2 - r_2)] &= x_1E[\alpha(v - Ev)C] \\ &= x_1E(\alpha vC) - x_1EvE(\alpha C)\end{aligned}$$

Recall that the hypothesis of decreasing absolute risk aversion implies that α is a decreasing function of π, and hence an increasing function of v: $\alpha = \alpha(v)$, with $\alpha'(v) > 0$. Let v_0 be the value of v at which $C = 0$, and note from the discussion in Section I that $v_0 > Ev$.

We may write the first term in (A6) as

$$\text{(A7)} \quad \Sigma \equiv x_1 \int_0^{v_0} v\alpha(v)C\phi(v)dv + x_1 \int_{v_0}^{\infty} v\alpha(v)C\phi(v)dv$$

where $\phi(v)$ is the probability density of v.

The first integral in (A7) is negative, and the second integral is positive, so we have

$$x_1 \int_0^{v_0} v\alpha(v)C\phi(v)dv > x_1v_0 \cdot \int_0^{v_0} \alpha(v)C\phi(v)dv$$

and $$x_1 \int_{v_0}^{\infty} v\alpha(v)C\phi(v)dv > x_1v_0 \cdot \int_{v_0}^{\infty} \alpha(v)C\phi(v)dv$$

hence $$\Sigma > x_1v_0 \int_0^{\infty} \alpha(v)C\phi(v)dv$$

It follows at once that (A6) is greater than $x_1(v_0 - Ev)E(\alpha C)$. But $(v_0 - Ev)$ is positive;

and since $EC = 0$, it follows that if α is increasing in v then $E(\alpha C) > 0$. Hence (A6) is greater than zero.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk-Bearing*, Chicago 1971.

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

H. E. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

M. B. Stewart, "Uncertainty and the Economics of Transfer Pricing," unpublished paper, Rutgers College 1977.

On the Comparative Statics of a Competitive Industry with Inframarginal Firms

By John C. Panzar and Robert D. Willig*

Recently, economists have begun to develop a theory of the perfectly competitive firm and industry in long-run equilibrium.[1] In contrast to the traditional model[2] in which *all* prices are parametric, this new theory takes explicit recognition of the fact that output price must adjust to exogenous changes in input prices before the industry can be said to be in long-run competitive equilibrium. The focus of this analysis has been to derive implications of competitive theory which can be tested using data generated by observing individual firms in long-run equilibrium.[3] This being the case, it was natural to assume (see, for example, Silberberg, p. 734) a horizontal industry supply curve, that is, that all firms are identical, and in equilibrium earn no rents. This assumption allowed the analysis to be carried out entirely at the level of the individual firm, since such firms behave in the long run "as if" they were minimizing average cost.

Unfortunately, this approach has also limited the applicability of the analysis at the *industry* level. The absence, by assumption, of inframarginal firms makes the theory inapplicable to a wide and important class of competitive industries with rising supply curves. In this note, we develop a simple analytical model of a competitive industry with a rising supply curve by positing that firms may have diverse endowments of a fixed factor in inelastic supply.[4] Thus in equilibrium there may be inframarginal firms earning economic rents (which are, of course, imputed to the fixed factor). Within this framework, we find that Ferguson and Saving's proposition that "In competitive industry, long-run equilibrium price will always vary directly with factor price, while long-run equilibrium output will vary inversely" (p. 780), is in general true only for normal factors of production.

Our results also have immediate applicability to the recent controversy[5] over whether or not the social benefits of an exogenous input price reduction can be equivalently evaluated using consumer's surplus measures from *either* the final product market *or* the relevant input market. This literature has also focused on the case of a competitive industry with a horizontal supply curve. We demonstrate that, when there are inframarginal firms, only the input market measure is accurate.

I. The Analytical Framework

We choose to represent the productive technologies of the firms by means of their minimum cost functions and their cost-minimizing input demand functions. We assume that the underlying production functions exhibit sufficient smoothness so that the former are twice continuously differentiable, and the latter are continuously differentiable.[6] In order to allow for firm

*Bell Laboratories, and Bell Laboratories and Princeton University, respectively. The views expressed are our own and do not necessarily reflect those of Bell Laboratories or the Bell System. We would like to thank George Borts and an anonymous referee for helpful comments and suggestions.

[1] The seminal paper in this literature is that of C. E. Ferguson and Thomas Saving.

[2] As developed, for example, by Paul Samuelson.

[3] See Eugene Silberberg for the most comprehensive development of this long-run theory. He also provides a valuable bibliography.

[4] Thus our model is an extension of what Jacob Viner, in his classic paper, referred to as the case of Ricardian increasing costs.

[5] This literature was initiated by Richard Schmalensee (1971). See also James Anderson, Daniel Wisecarver, and Schmalensee (1976).

[6] It suffices that the production functions are twice continuously differentiable and strongly quasi concave (see Silberberg).

diversity, we assume that these functions are indexed by a (nonnegative) cost decreasing parameter θ. Without loss of generality, θ may be viewed as indexing a firm's endowment of a fixed factor in inelastic supply, such as a firm's location or the nitrogen content of a farm's soil. For simplicity, we assume that this factor has no alternative uses outside the industry in question.

A firm of type θ is thus characterized by the cost and input demand functions,

$$(1)\quad C(q,w,\theta),\quad \frac{\partial C}{\partial \theta} \equiv C_\theta < 0, \quad \frac{\partial C}{\partial q} \equiv C_q > 0,\quad \frac{\partial C}{\partial w_j} \equiv C_{w_j} = x_j^d(q,w,\theta) \geq 0, \quad j = 1,\dots,n$$

where q is output, $w = (w_1, \dots, w_n)$ is a strictly positive vector of input prices, and x_j^d represents the cost-minimizing level of employment of the jth input. (The last equality in (1) is a result of the well-known Shephard's Lemma.)

The equilibrium of the price-taking competitive firm of type θ is given by the solution of the program,

$$\max_{q \geq 0}\ [pq - C(q,w,\theta)]$$

which has as necessary and sufficient conditions:

$$(2)\quad p - C_q(q,w,\theta) \leq 0, \quad q \geq 0, \quad q(p - C_q) = 0$$

$$(3)\quad C_{qq} \geq 0$$

We will assume that these conditions are satisfied with $q > 0$ and $C_{qq} > 0$. Equations (2) and (3) can now be used to derive the supply curve of the individual firm, $q^*(p,w,\theta)$. This function has the following partial derivatives:

$$(4)\quad \frac{\partial q^*}{\partial p} = \frac{1}{C_{qq}} > 0, \quad \frac{\partial q^*}{\partial w_j} = -\frac{\partial x_j^d}{\partial q}\Big/ C_{qq}, \quad \frac{\partial q^*}{\partial \theta} = -C_{q\theta}/C_{qq}$$

II. Industry Equilibrium

Crucial to the characterization of industry equilibrium is the notion of ***marginal firms***; that is, those firms whose θ allows them to (optimally) earn exactly zero profits. This margin, $\hat{\theta}(p,w)$, is defined by the following equation:

$$(5)\quad pq^*(p,w,\hat{\theta}) - C[q^*(p,w,\hat{\theta}),w,\hat{\theta}] = 0$$

Since (maximized) profits are obviously an increasing function of the cost-reducing parameter θ, firms with θ less than $\hat{\theta}$ will not participate in the industry. Making use of (2) and (5), we can see that $\hat{\theta}(p,w)$ has the following properties:

$$(6)\quad \frac{\partial \hat{\theta}}{\partial p} = -\left[(p - C_q)\frac{\partial q^*}{\partial p} + q^*\right] \div \left[(p - C_q)\frac{\partial q^*}{\partial \theta} - C_\theta\right] = \frac{q^*}{C_\theta} < 0$$

$$(7)\quad \frac{\partial \hat{\theta}}{\partial w_j} = -\left[(p - C_q)\frac{\partial q^*}{\partial w_j} - C_{w_j}\right] \div \left[(p - C_q)\frac{\partial q^*}{\partial \theta} - C_\theta\right] = -\frac{x_j^d}{C_\theta} \geq 0$$

Equations (6) and (7) indicate that the minimum value of θ which is required for firms to participate in the industry is, as one would expect, a decreasing function of output price and an increasing function of all input prices.

We can now write the industry supply curve as

$$(8)\quad S(p,w) = \int_{\hat{\theta}(p,w)}^{\bar{\theta}} q^*(p,w,\theta) f(\theta)\, d\theta$$

where $f(\theta) \geq 0$ is the (exogenous) density function which indicates how many firms there are of type θ ($\bar{\theta}$ is the maximum value of θ observed in the population). Industry equilibrium occurs at that price p^e, where industry supply equals industry demand. That is,

$$(9)\quad S(p^e,w) - D(p^e) = 0$$

where $D(p)$ is the (downward-sloping) demand curve for the output of the industry.

III. Comparative Statics

In order to determine the effect of an input price change on the equilibrium output price, we substitute (8) into (9), totally differentiate, and obtain

$$(10)\quad \frac{\partial p^e}{\partial w_i} = \frac{-\dfrac{\partial S}{\partial w_i}}{\dfrac{\partial S}{\partial p} - D'} = -\left[\int_{\hat\theta}^{\bar\theta} \frac{\partial q^*}{\partial w_i} f(\theta)d\theta - q^*(p^e, w, \hat\theta) f(\hat\theta)\frac{\partial \hat\theta}{\partial w_i}\right] \div \left[\int_{\hat\theta}^{\bar\theta} \frac{\partial q^*}{\partial p} f(\theta)d\theta - q^*(p^e, w, \hat\theta) f(\hat\theta)\frac{\partial \hat\theta}{\partial p} - D'(p^e)\right]$$

While (4) and (6) guarantee that the denominator of (10) is positive, the sign of the numerator is in general indeterminate. The following (standard) definitions will facilitate interpretation of the result.

DEFINITION: An input x_i is said to be (locally)[7] *inferior* if $\partial x_i^d(\cdot)/\partial q < 0$, $\forall\ \theta$. Otherwise, it is said to be (locally) normal.

DEFINITION: An input x_i is said to be (locally) *strongly* normal if $\partial x_i^d(\cdot)/\partial q > 0$, $\forall\ \theta$.

PROPOSITION 1: *If x_i is (locally) strongly normal, then $\partial p^e/\partial w_i > 0$.*

PROOF:

Since $\partial S/\partial p - D' > 0$, it is clear that

$$(11)\quad sgn\left(\frac{\partial p^e}{\partial w_i}\right) = -\,sgn\left(\frac{\partial S}{\partial w_i}\right)$$

From (4) and (10), we see that

$$(12)\quad \frac{\partial S}{\partial w_i} = \int_{\hat\theta}^{\bar\theta} \left[-\frac{\partial x^d}{\partial q}\Big/ C_{qq}\right] f(\theta)d\theta - q^*(p^e, w, \hat\theta) f(\hat\theta)\frac{\partial \hat\theta}{\partial w_i}$$

From (7) and strong normality, it is clear that $\partial S/\partial w_i < 0$, and (11) yields the result.

Proposition 1 differs from the result derived by Ferguson and Saving for the case of identical marginal firms; that is, that the equilibrium output price is a strictly increasing function of all input prices *regardless* of the normality properties of the inputs. That result is hardly surprising since, with free entry, the industry supply curve is perfectly elastic at the level of the minimum point of average cost. An increase in the price of any input raises that level, and consequently the equilibrium output price must rise.

In the present framework, the shift in the industry supply curve resulting from a factor price increase has two components; the change in the output levels of *inframarginal* firms, and the exit of marginal firms (and their output) from the industry. When the input in question is normal, these two effects operate in the same direction. However, in the case of inferior factors, the effects oppose each other, and the sign of the total effect cannot be resolved without reference to the relative numbers of marginal vs. inframarginal firms; it cannot be resolved from technological considerations alone. The following proposition states this result positively:

PROPOSITION 2: *If x is (locally) inferior, and if all firms are inframarginal at the initial equilibrium* (*i.e.*, $f[\hat\theta(p^e, w)] = 0$), *then* $\partial p^e/\partial w_i < 0$.

PROOF:

Under these conditions, (12) becomes

$$\frac{\partial S}{\partial w_i} = \int_{\hat\theta}^{\bar\theta} \left[-\frac{\partial x_i^d}{\partial q}\Big/ C_{qq}\right] f(\theta)d\theta > 0$$

and (11) yields the result.

[7] The term "locally" refers only to the arguments *other than θ*. This form of definition is required because the normality properties of an input may vary with θ unless further assumptions are made.

Intuitively, when all firms are earning positive rents, a "small" increase in a factor price will not lead to the exit of any firms, although all will have lower profits. Thus the expansion of each firm's output resulting from an increase in the price of an inferior factor is directly translated into an outward shift of the industry supply curve, resulting in a lower equilibrium price.

It is important to point out that the competitive equilibrium which we analyze is well defined even when there are no marginal firms. Equation (5) merely serves to identify the level of θ below which market participation is unprofitable. Equation (8), which characterizes industry equilibrium, remains perfectly well defined even if $f(\theta) = 0$ over wide ranges. Notice also that Proposition 2 remains valid even if all (inframarginal) firms are identical.[8] Thus it is *not* the assumption of identical firms which drives the results of the new long-run theory of the firm, but rather the assumption that *all* firms are *marginal*.

IV. On the Choice of Welfare Measures

Recently, there has arisen a literature on the measurement of the social benefits resulting from an input price change. Under the assumption of identical perfectly competitive firms, all of which are marginal, welfare effects can be *equivalently* measured as the change in consumer's surplus in either the (relevant) input market *or* the final product market (see Anderson).

We show that this equivalence vanishes when there are inframarginal firms. We begin by demonstrating, in the context of the present model, that welfare changes can be measured as changes in the area under the industry factor demand curve.

PROPOSITION 3: *The partial derivative of welfare with respect to the jth input price is exactly equal to the negative of industry demand for that input.*

PROOF:

We take as our welfare measure the sum of producer's and consumer's surpluses in this industry.[9] Thus

$$W = \int_{p^e(w)}^{\infty} D(p)dp + \int_{\theta[p^e(w),w]}^{\bar{\theta}} \{p^e(w)q^*[p^e(w),w,\theta] - C(q^*,w,\theta)\} f(\theta)d\theta \tag{13}$$

Differentiating (13) with respect to the jth input price yields

$$\begin{aligned} \frac{\partial W}{\partial w_j} = &- D(p^e)\frac{\partial p^e}{\partial w_j} \\ &+ \int_{\hat{\theta}}^{\bar{\theta}} \left\{ (p^e - C_q)\left(\frac{\partial q^*}{\partial p}\frac{\partial p^e}{\partial w_j} + \frac{\partial q^*}{\partial w_j}\right) + q^*\frac{\partial p^e}{\partial w_j} - C_{w_j}\right\} f(\theta)d\theta \\ &- [p^e(w)q^*[p^e(w),w,\hat{\theta}] - C(q^*,w,\hat{\theta})]f(\hat{\theta})\left[\frac{\partial\hat{\theta}}{\partial p}\frac{\partial p^e}{\partial w_j} + \frac{\partial\hat{\theta}}{\partial w_j}\right] \end{aligned} \tag{14}$$

Upon substituting (1), (2), (5), (8), and (9) into (14), we have

$$\frac{\partial W}{\partial w_j} = -\int_{\hat{\theta}}^{\bar{\theta}} x_j^d(q^*,w,\theta)f(\theta)d\theta \equiv -X_j(w) \tag{15}$$

where $X_j(w)$ is the industry factor demand function.[10]

[8] Since all integrals used in the analysis can be viewed as Stieltjes integrals, $f(\theta)$ need not be continuous and may in fact have all its mass at a single point.

[9] We shall not be concerned with the effects on the industry actually producing the input. One might assume that the input is imported or that it is produced by a competitive industry with a perfectly elastic supply curve. In such cases a change in input price has no subsidiary effects on (domestic) welfare.

[10] Notice that this industry input demand curve $X_j(w)$ is of the "reduced form" variety because all *endogenous* variables, such as output price and the number of firms, have been eliminated as arguments.

It is clear from (15) that an input price increase *always* reduces welfare. To calculate the welfare cost of any finite increase, it is necessary to integrate (15); thus the welfare change can be measured as the area under the industry's input demand curve.[11] All that is required to demonstrate that this change cannot in general be equivalently measured as the change in consumer's surplus in the output market is to recall Proposition 2, which states that the equilibrium output price may actually fall when the price of an inferior input increases. In such a case, the surplus of final consumers would actually increase, but we know, via (15), that total welfare must decrease. In general, the change in the surplus of final consumers will equal the total welfare change only when there is no profit effect; this can occur only when there are no inframarginal firms.[12]

V. Conclusion

A perfectly competitive industry may exhibit a rising supply curve and inframarginal firms, even in long-run equilibrium. The comparative statics behavior of such industries may be qualitatively quite different from that of competitive industries with horizontal supply curves. In particular, when there are inframarginal firms, we have shown that an increase in the price of a factor of production may actually result in a *reduction* in the equilibrium output price. A direct implication of this result is that the change in consumer's surplus in the output market will not, in general, accurately measure the welfare effects of a change in input prices.

[11] Equation (15) and its subsequent integration can also be derived from a model with identical marginal firms, thereby allowing welfare changes to be measured in the input market in that case as well

[12] Note that we have shown that the equivalence does not even hold locally Thus the discrepancy is *not* due to the inequality of consumer's surplus and the compensating variation for finite changes, an issue discussed by Anderson As is well known, the derivatives of the two measures are equal

REFERENCES

J. E. Anderson, "The Social Cost of Input Distortions: A Comment and a Generalization," *Amer. Econ. Rev.*, Mar. 1976, *66*, 235-38.

C. E. Ferguson and T. R. Saving, "Long-Run Scale Adjustments of a Perfectly Competitive Firm and Industry," *Amer. Econ. Rev.*, Dec. 1969, *59*, 774-83.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.*, Sept. 1971, *61*, 682-87.

———, "Another Look at the Social Valuation of Input Price Changes," *Amer. Econ. Rev.*, Mar. 1976, *66*, 239-43.

E. Silberberg, "The Theory of the Firm in 'Long-Run' Equilibrium," *Amer. Econ. Rev.*, Sept. 1974, *64*, 734-41.

J. Viner, "Cost Curves and Supply Curves," *Z. Nationalökon.*, Sept. 1931, *3*, 23-46; reprinted in Kenneth Boulding and George Stigler, eds., *Readings in Price Theory*, Chicago 1952.

D. Wisecarver, "The Social Cost of Input-Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359-71.

NOTES

NINETY-FIRST ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

Chicago, Illinois, August 28-31, 1978

Preliminary Announcement of the Program

Monday, August 28, 1978

10:00 A.M. EXECUTIVE COMMITTEE MEETING

Tuesday, August 29, 1978

8:00 A.M. NEW DIRECTIONS FOR EMPLOYMENT POLICY*
Presiding ISABEL SAWHILL, National Commission for Manpower Policy
Papers. GEORGE E. JOHNSON AND ARTHUR E. BLAKEMORE, Council of Economic Advisors
The Potential Efficacy of Employment Policy in Reducing the Noninflationary Unemployment Rate
JOHN BISHOP AND ROBERT HAVEMAN, University of Wisconsin
Selective Employment Subsidies. Can Okun's Law Be Repealed?
RONALD G. EHRENBERG, Cornell University
The Impact of Retirement Policies on Employment and Unemployment
Discussants DANIEL S HAMERMESH, Michigan State University
KATHLEEN P CLASSEN, Center for Naval Analyses
CORDELIA REIMERS, Princeton University

8:00 A.M. INTERNATIONAL COMMODITY MARKETS AND AGREEMENTS*
Presiding JERE BEHRMAN, University of Pennsylvania
Papers ROBERT PINDYCK, Massachusetts Institute of Technology
Cartelization of World Commodity Markets
DAVID BIGMAN, International Monetary Fund, AND SHLOMO REUTLINGER, World Bank
National and International Policies toward Food Security and Price Stabilization
F G ADAMS AND J BEHRMAN, University of Pennsylvania
Impact of Commodity Stabilization on Economic Development Some Empirical Studies
BRUCE GARDNER, Texas A&M University
Optimal Inventory Control and Grain Price Stabilization
Discussants PHILIP ABBOTT, Northeastern University
WALTER LABYS, West Virginia University

8:00 A.M. INCREASING THE VIABILITY OF CENTRAL CITIES NEW STRATEGIES, OLD STRATEGIES (Joint Session with the National Economics Association)*
Presiding KARL D. GREGORY, Oakland University
Papers CLEVELAND A CHANDLER AND WILFRED L DAVID, Howard University
Alternative Policies for the Revitalization of Central Cities
CHARLES E. ANDERSON, Livingston College, Rutgers University
Production of Health Services Can Hospital Costs be Reduced?
SAMUEL L. MYERS, JR., University of Texas-Austin, AND KENNETH E PHILLIPS, Rand Corporation
Housing Segregation and Black Employment Another Look at the Ghetto Dispersal Strategy
Discussants. BENNETT HARRISON, Massachusetts Institute of Technology
MARTHA BLAXALL, Department of Health, Education, and Welfare
BRUCE DUNSON, Harvard University

8:00 A.M. TOPICS IN FISCAL POLICY
Presiding. RICHARD F KOSOBUD, University of Illinois at Chicago Circle
Papers. JAMES BARTH AND JOSEPH CORDES, George Washington University
Substitutability, Complementarity and the Impact of Government Spending on Economic Activity

GERALD A. CARLINO, University of Missouri-Kansas City
Employing a Variable Sales Tax Mechanism for Stabilization Purposes
DONALD A. HANSON AND RAMA RAMACHANDRAN, Southern Methodist University
Public Choice and Debt Financing
L. DOUGLAS LEE, Joint Economic Committee
The Budget Control Act of 1974 and Fiscal Policy
Discussants. RICHARD F. KOSOBUD, University of Illinois at Chicago Circle
HERMAN I. LIEBLING, Lafayette College

8:00 A.M. GOVERNMENT POLICY AND INNOVATION IN HEALTH CARE
Presiding: CAROLE KITTI, National Science Foundation
Papers. LOUISE RUSSELL, The Brookings Institution
Determinants of Technological Diffusion in the Hospital Sector
JUDITH WAGNER, The Urban Institute
Problems in the Measurement of Costs and Benefits of New Medical Technologies
RONALD HANSEN, University of Rochester
The Impact of Public Policy on Pharmaceutical Innovation
Discussants ANTHONY ROMEO, University of Connecticut
MICHAEL RADISCH, General Accounting Office
ROBERT HELMS, American Enterprise Institute

8:00 A.M. BEHAVIOR OF REGULATED FIRMS
Presiding RAJINDAR K. KOSHAL, Ohio University
Papers: DOUGLAS W. CAVES, LAURITS R. CHRISTENSEN, University of Wisconsin-Madison, AND JOSEPH A. SWANSON, Northwestern University
Economic Performance in Regulated and Unregulated Environments: A Case Study of the U.S. and Canadian Railroad Industries
FERDINAND K. LEVY, Georgia Institute of Technology, AND GLORIA M. SHATTO, Trinity University
Electrical Utilities: Philanthropy, Regulation, and the Rate of Return
GEORGE SWEENEY, Vanderbilt University
Adoption of Cost Saving Innovations by a Regulated Firm
Discussants JAMES T. H. TSAO, AT&T
MILES H. SONSTEGAARD, University of Arkansas

8:00 A.M. FINANCIAL INTERMEDIATION
Presiding: M. A. AKHTAR, Federal Reserve Bank of New York
Papers E. TYLOR CLAGETT AND STANLEY R. STANSELL, University of Houston
Scale Economies in a Cooperative Financial System
GILLIAN GARCIA, University of California-Berkeley
The U.S. Currency Ratio
JANG H. YOO AND STEVEN M. CRAFTON, Virginia Commonwealth University
The Impact of FNMA on The Mortgage Market
LAWRENCE S. DAVIDSON, JOHN M. FINKELSTEIN, AND RAMESH K. S. RAO, Indiana University
Optimal Intermediation: A Synthesis
JOHN M. MASON, Wharton School
A Theory of the Firm, Finance, and Financial Intermediation
Discussant M. A. AKHTAR, Federal Reserve Bank of New York

8:00 A.M. STATE AND LOCAL FINANCE
Presiding CHARLES WALDAUER, Widener College
Papers. ELEANOR D. CRAIG, University of Delaware, AND A. JAMES HEINS, University of Illinois
The Effect of Tax Elasticity on Government Spending
SAMUEL H. BAKER, College of William and Mary
An Empirical Test of the Campbells' Hypothesis
IRA EPSTEIN, Maryland-National Capital Park and Planning Commission
A Property Tax Base Model
Discussants SHLOMO MAITAL, Princeton University
CHARLES WALDAUER, Widener College

8:00 A.M. REGULATING THE OPERATING POLICIES OF BANK-HOLDING COMPANIES (Joint Session with the American Finance Association)
Presiding: ROBERT A. EISENBEIS, Federal Reserve System

Papers. CHUN LAM, Tulane University
Bank-Holding Company Behavior and the Effectiveness of Capital Regulation
LUCILLE S. MAYNE, Case Western Reserve University
Bank-Holding Company Characteristics and the Upstreaming of Bank Funds
DONALD T SAVAGE, Federal Reserve System
Bank-Holding Company Management and Service Fees
Discussants. ARNOLD HEGGESTAD, University of Florida
JOHN PRINGLE, University of North Carolina
WILLIAM A LONGBRAKE, Federal Deposit Insurance Corporation

8 00 A.M TOPICS IN ECONOMIC THEORY
Presiding A L LEVINE, University of New Brunswick
Papers J. BARKLEY ROSSER, JR AND RAYMOND PRINCE, James Madison University
Reswitching between Food and Energy Production in the Western United States
MURRAY BROWN AND MARILYN MANSER, State University of New York-Buffalo
Marriage, Commodity Demands, and Labor Supply Decisions in the Aggregate
STEVEN SHAVELL, Harvard University
Accidents, Liability and Insurance
Discussants MICHAEL SATTINGER, Miami University, Ohio
WALTER J WESSELS, North Carolina State University at Raleigh
A L LEVINE, University of New Brunswick

10 15 A M APPLIED WELFARE THEORY*
Presiding E.E. BAILEY, Civil Aeronautics Board
Papers JEFFREY HARRIS, Massachusetts Institute of Technology
Pricing Rules for Hospitals
J C PANZAR, Bell Laboratories
Equilibrium and Welfare in Unregulated Airline Markets
R.D WILLIG, Bell Laboratories and Princeton University, AND E E BAILEY, Civil Aeronautics Board
The Economic Gradient Method Theory and Application to Postal and Telephone Services
Discussants M PAULY, Northwestern University
I KELLER, University of California-Berkeley
W VICKREY, Columbia University

10 15 A M APPRAISING THE NATION'S LABOR FORCE STATISTICS (Joint Session with the Industrial Relations Research Association)
Presiding CHARLES C KILLINGSWORTH, Michigan State University
Papers ARVIL V ADAMS, National Commission on Employment and Unemployment Statistics (NCEUS)
Who's in the Labor Force A Simple Counting Problem?
DIANE WERNEKE, NCEUS
Measuring Economic Hardship in the Labor Market
CURTIS GILROY, NCEUS
Counting the Labor Force with the Current Population Survey
Discussant· STANLEY RUTTENBERG

10 15 A.M. CURRENT ISSUES IN EAST-WEST TRADE AND PAYMENTS (Joint Session with Association for Comparative Economic Studies)*
Presiding. ARCADIUS KAHAN, University of Chicago
Papers. PADMA DESAI, Russian Research Center (Harvard) and Boston University
The Productivity of Foreign Credits to the Soviet Economy
FRANKLYN HOLZMAN, Russian Research Center (Harvard) and Tufts University
A Theory of the Persistent Hard-Currency Shortages of Centrally Planned Economies
STEVEN ROSEFIELDE, University of North Carolina
East-West Trade and Postwar Soviet Economic Growth. A Sectoral Production Function Approach
Discussants· EDWARD HEWETT, University of Texas
MICHAEL DOLAN, Queens College

10 15 A M BEHAVIOR UNDER UNCERTAINTY
Presiding JAMES R. MEGINNISS, Claremont Graduate School
Papers MATTHEW BLACK, Mathematica Policy Research
Empirical Test of the Impact of Future Income Uncertainty on Saving Decisions

JACOB PAROUSH AND NAVA KAHANA, Bar-Ilan University
Price Uncertainty and the Cooperative Firm
STEPHEN MCCAFFERTY, Ohio State University
Optimal Wage and Layoff Policies in Stochastic Equilibrium
PETER LINNEMAN, University of Chicago
An Empirical Analysis of Behavior in Markets with Risky Products

10.15 A.M. AGING AND RETIREMENT
Presiding JOSEPH W. HUNT, Shippensburg State College
Papers. ROBERT CLARK, North Carolina State University, AND JOSEPH J. SPENGLER, Duke University
Economic Implications of Population Aging
RICHARD V. BURKHAUSER, University of Wisconsin, AND JOHN A. TURNER, Social Security Administration
Life Cycle Welfare Costs of Social Security
THAD W MIRER, State University of New York-Albany
The Wealth-Holding Behavior of the Aged
Discussant JOSEPH W. HUNT, Shippensburg State College

10.15 A.M. ECONOMIC ANALYSIS IN AN LDC CONTEXT
Presiding. BANKEY L. SHARMA, Jackson State College
Papers. MAXWELL J. FRY, Bogazici University
Turkey's Upward-Sloping Phillips Curve
PREM S LAUMAS, Northern Illinois University
A Cross-Section Study of Business Demand for Liquid Assets in an Underdeveloped Economy
KATRINE W. SAITO, World Bank, AND PARTHASARATHI SHOME, American University
The Impact of Social Security Institutions on Resource Mobilization and Allocation The Asian Experience
C. MICHAEL HENRY, New Brunswick, NJ
Economics of Adoption of New Farm Technology The Case of The Guyanese Rice Industry
Discussants MARTIN SCHNITZER, Virginia Polytechnic Institute and State University
BANKEY L. SHARMA, Jackson State College
RICHARD S THORN, University of Pittsburgh

10.15 A.M. THE ECONOMIC ANALYSES OF HOSPITALS (Joint Session with Health Economics Research Organization)
Presiding. DONALD E. YETT, University of Southern California
Papers: GESTUR B. DAVIDSON, University of Minnesota
Manpower Substitution and Hospital Efficiency
FRANK A. SLOAN AND BRUCE STEINWALD, Vanderbilt University
Effects of Regulation on Hospital Demand for Inputs
JUDITH D. BENTKOVER, Arthur D. Little, Inc
The Impact of Inflation and Unemployment on Public Hospital Utilization
Discussants LEONARD DRABEK, Bureau of Health Manpower, HEW
RICHARD SCHEFFLER, Institute of Medicine, National Academy of Sciences
JOHN RAFFERTY, National Center for Health Service Research

10.15 A.M DISEQUILIBRIUM ESTIMATION (Joint Session with the Econometric Society)
Presiding RICHARD QUANDT, Princeton University
Papers: TAKATOSHI ITO, Harvard University
Methods of Estimation for Disequilibrium Macro-Economic Models
MARK GERSOVITZ, Princeton University
Estimation of Systems with Inequalities The Two-Gap Models
Discussant: RICHARD QUANDT, Princeton University

2:00 P.M SELF-SELECTED AND CENSORED SAMPLES (Joint Session with the Econometric Society)
Presiding: JAMES J HECKMAN, University of Chicago
Papers. (To be announced)

2:00 P.M URBAN ECONOMIC ANALYSIS IN LESS DEVELOPED COUNTIRES
Presiding: JOHN R MEYER, Harvard University
Papers: GREGORY K. INGRAM, Harvard University and World Bank
The Spatial Structure of Latin American Cities

KYU SIK LEE, World Bank
Intra-Urban Location of Manufacturing Employment in Colombia
EDWIN S. MILLS, Princeton University
A Comparison of Urban Density Patterns in Japan, Korea, and the United States
WILLIAM C. WHEATON, Massachusetts Institute of Technology
Housing Policies and Urban Markets in LDCs. Egypt
Discussants. DENNIS CARLTON, University of Chicago
ROGER SCHMENNER, Harvard University
BERTRAND RENAUD, The World Bank

2.00 P.M. THE EFFECTIVENESS OF FISCAL POLICY*
Presiding: ALAN S. BLINDER, Princeton University
Papers. CARL F. CHRIST, Johns Hopkins University
Implications of the Government Budget Constraint
WALTER C. DOLDE, JR., Carnegie-Mellon University
Temporary Taxes as Macro-Economic Stabilizers
RAY C FAIR, Yale University
On Modeling the Effects of Government Policies
Discussants WILLEM BUITER, Princeton University
MICHAEL DARBY, University of California-Los Angeles
THOMAS SARGENT, University of Minnesota

2 00 P M PRICING IN REGULATED INDUSTRIES
Presiding FRIEDA REITMAN, University of Connecticut
Papers· JAN PAUL ACTON AND WILLARD MANNING, Rand Corporation
Residential Energy Demand Under Time-Of-Day Pricing
SCOTT ATKINSON, U.S Department of Energy
A Comparative Analysis of Two Electricity Time-Of-Day Pricing Experiments
GEORGE E. HOFFER, Virginia Commonwealth University
Pricing Motor Vehicle Liability Insurance at the Margin
A H STUDENMUND, Occidental College and the Transportation Systems Center (USDOT)
A Preliminary Analysis of the Fare-Free Transit Experiment
Discussant FRIEDA REITMAN, University of Connecticut

2 00 P M ADVERTISING AND CONSUMER SOVEREIGNTY
Presiding ALLEN R FERGUSON, Public Interest Economics Foundation
Papers RANDALL BARTLETT, Federal Trade Commission
Marketing Ideologies Influencing Social Values
KELVIN LANCASTER, Columbia University
Advertising and Consumer Choice
ALI M REZA, University of Pittsburgh
Advertising's Effect on the Marginal Propensity to Consume Out of Permanent Income
Discussants ZENA COOK, Public Interest Economics Foundation
BOBBY JOE CALDER, School of Business, Northwestern University

2.00 P M. COMPETITION AND REGULATION IN THE HEALTH CARE SECTOR (Joint Session with the Health Economics Research Organization)
Presiding EDWARD S MILLS, Blue Cross Association and Blue Shield Association
Papers JOSEPH NEWHOUSE, The Rand Corporation
The Erosion of the Medical Market Place
WILLIAM LYNK, Blue Cross Association and Blue Shield Association
Market Organization and the Level of Costs A Study of Blue Cross Blue Shield Administrative Expenses
KEITH LEFFLER, University of Rochester
Physician Licensure. Competition and Monopoly in American Medicine
Discussants MARK PAULY, Northwestern University
WARREN GREENBERG, Federal Trade Commission
RICHARD ARNOULD, University of Illinois

2·00 P.M. WAGES AND EMPLOYMENT*
Presiding EDMUND S PHELPS, New York University
Papers· GUILLERMO A CALVO, Columbia University
Wage Theory and Unemployment

JOHN B. TAYLOR, Columbia University
Staggered Wage Setting in a Macro Model
OLIVIER J. BLANCHARD, Harvard University
Backward and Forward Solutions for Economies with Rational Expectations
Discussants. (To be announced)

2:00 P.M. SELECTING ENERGY STRATEGIES: THEIR IMPACT ON THE DISADVANTAGED (Joint Session with the National Economic Association)
Presiding. FLOURNOY A. COLES, JR., Vanderbilt University
Papers. EVERSON A. W. HULL, American Petroleum Institute
Implications of U.S. Energy Policy and A Vibrant Economy for Employment Opportunities of Black Americans
BERNARD E. ANDERSON, University of Pennsylvania
Energy Policy and Black Employment Projection: A Preliminary Analysis
LENNEAL HENDERSON, Joint Center for Political Studies
Public Utilities and the Poor: The Socioeconomic Impact of Reform Proposals
JOHN M. BRAZZEL AND LEON J. HUNTER, Department of Energy
A Distributional Analysis of Trends in Energy Expenditures by Black Households
ERNEST WILSON III, University of Pennsylvania
The Determinants of Production and Distribution Decisions in the Energy Sector of Nigeria and Zaire
Discussants. WENDELL BUTLER, Department of Energy
CHARLES BEISLY, Congressional Budget Office
RICHARD MORGENSTERN, Congressional Budget Office
SHIRLEY BURGGRAF, Florida A&M University

2:00 P.M. WHAT ECONOMISTS THINK*
Presiding. CAROLYN BELL, Wellesley College
Paper. J. KEARL, C. POPE, C. WHITING, AND L. WIMMER, Brigham Young University
What Do Economists Agree On?
Discussants. CAROLYN BELL, Wellesley College
LEONARD SILK, *New York Times*
BRUCE LIPPKE, Weyerhaeuser Company

2:00 P.M. MORE ON KEYNES (Joint Session with the History of Economics Society)
Presiding. WILLIAM O. THWEATT, Vanderbilt University
Papers. ROBERT CLOWER, University of California-Los Angeles
The End of Keynesian Economics
DON PATINKIN, Hebrew University
New Materials on the Development of Keynes' Monetary Thought
Discussants. THOMAS K. RYMES, Carleton University
DON MOGGRIDGE, University of Toronto

2:00 P.M. RECENT DEVELOPMENTS IN THE DEMAND FOR MONEY*
Presiding. THOMAS MAYER, University of California-Davis
Papers. THOMAS CARGILL, University of Nevada-Reno, RENE MEYER, AND ROBERT MEYER, University of California-Berkeley
Stability of the Demand Function for Money: An Unresolved Issue
CHARLES LIEBERMAN, University of Maryland
Technological Change and Structural Change in the Demand for Money
GILLIAN C. GARCIA AND SIMON PAK, University of California-Berkeley
Some Clues in the Case of the Missing Money
Discussants. JOSEPH BISIGNANO, Federal Reserve Bank of San Francisco
MICHAEL HAMBURGER, Federal Reserve Bank of New York

2:00 P.M. FINANCIAL CONSEQUENCES OF INFLATION AND INFLATION RISK (Joint Session with the American Finance Association)
Presiding. DUDLEY G. LUCKETT, Iowa State University
Papers. JOHN H. MAKIN AND MAURICE D. LEVI, University of Washington
Fisher, Phillips, Friedman, and the Measured Impact of Inflation on Interest
KAJAL LAHIRI AND JUNG SOO LEE, State University of New York-Albany
Tests of Rational Expectations and the Fisher Effect
MARK J. FLANNERY, University of Pennsylvania
Interest Rate Variability and Financial-Intermediary Welfare

Discussants: ANTHONY M. SANTOMERO, University of Pennsylvania
ALEX CUKIERMAN, New York University
J. HUSTON MCCULLOCH, Boston College

8.00 P.M. RICHARD T. ELY LECTURE*
Presiding: ROBERT M. SOLOW, Massachusetts Institute of Technology
Speaker: ALFRED E. KAHN, Chairman, Civil Aeronautics Board
Applications of Economics to an Imperfect World

Wednesday, August 30, 1978

8.00 A.M. TOPICS IN ENVIRONMENTAL ECONOMICS
Presiding: J. T. BRIMER, Troy, Idaho
Papers: GARDNER BROWN, University of Washington
Valuation of Nonmarket Natural Resources
MALCOLM DOLE ET AL
Estimation of Vegetation Damage from Ambient Air Pollution
JOHN A. SORRENTINO, JR., Temple University
Coordinating the Effort to Abate Aircraft Noise

8.00 A.M. NEW DIRECTIONS IN INDUSTRIAL ORGANIZATION*
Presiding: SANFORD GROSSMAN, University of Pennsylvania
Papers: STEVEN SALOP, Civil Aeronautics Board
Strategic Entry Deterrence
JOSEPH STIGLITZ, Oxford University
Information and Market Structure
ROBERT WILLIG, Princeton University
Multiproduct Technology and Market Structure
Discussant: SANFORD GROSSMAN, University of Pennsylvania

8.00 A.M. DIFFUSION OF TECHOLOGY INTO AND OUT OF THE UNITED STATES
Presiding: SUMIYE OKUBO, National Science Foundation
Papers: EDWIN MANSFIELD, University of Pennsylvania
Studies of the Relationship between International Technology Transfer and R&D Expenditures by U.S. Firms
RAYMOND VERNON, Harvard University
The International Spread of Innovations via the Foreign Subsidiaries of U.S. Based Enterprises
W. HALDER FISHER, Battelle Memorial Institute
Technology Transfer as a Motive for U.S. Investment by Foreign Firms
Discussants: GARY HUFBAUER, U.S. Treasury Department
ROLF PIEKARZ, National Science Foundation
(To be announced)

8.00 A.M. PROSPECTS OF AN ECONOMIC CRISIS IN THE 1980'S
Presiding: (To be announced)
Papers: DANIEL FUSFELD, University of Michigan
The Next Great Depression
PAUL DAVIDSON, Rutgers University
Is Monetary Collapse in the 1980's in the Cards?
HELEN JUNZ, Department of the Treasury
Structural Changes and Adjustment Policies
Discussants: LEONARD SILK, *New York Times*
ALAN FERGUSON, Public Interest Economics Foundation

8.00 A.M. URBAN DEVELOPMENT: STRUCTURING FEDERAL POLICY
Presiding: KATHARINE C. LYALL, Department of Housing and Urban Development
Papers: GEORGE PETERSON, The Urban Institute
The Condition of Public Infrastructure
DAVID PURYEAR, The Maxwell School, Syracuse University
Urban Development: Capital Subsidies vs. Wage Subsidies
ELIZABETH ROISTACHER, Queens College
Employment, Unemployment, and Migration: A Microeconomic Analysis, 1968-1976.

ROGER SCHMENNER, Harvard Business School
The Location Decision of Firms
Discussants CHARLES LEVEN, Washington University
BERNARD ANDERSON, University of Pennsylvania
RAY STRUYK, Department of Housing and Urban Development

10 15 A M CONTROLLING INFLATION INCENTIVES FOR WAGE AND PRICE STABILITY*
Presiding HENRY WALLICH, Federal Reserve Board
Papers LAWRENCE SEIDMAN, University of Pennsylvania
The Role and Design of a Tax Based Incomes Policy
DONALD NICHOLS, U.S. Department of Labor
Similarities and Differences between TIP and Price Controls
RICHARD SLITOR, Economic Consultant
Implementation and Design of Tax Based Income Policies
Discussants. BARRY BOSWORTH, Council on Wage and Price Stability
WILLIAM VICKREY, Columbia University

10 15 A.M SOCIAL SECURITY ITS FINANCING AND FUTURE*
Presiding ALICIA MUNNELL, Federal Reserve Bank of Boston
Papers A. HAEWORTH ROBERTSON, Social Security Administration
The Financing and Future of the Social Security Program (OASDHI)
PAUL VAN DER WATER, Department of Health, Education, and Welfare
Disability Insurance (DI)
UWE F REINHARDT, Princeton University
Medicare (HI)
Anthony Pellechio, Harvard University
Old Age and Survivors Insurance (OASI)
Discussants JOSEPH A PECHMAN, The Brookings Institution
MICHAEL K TAUSSIG, Rutgers The State University

10 15 A M THE ECONOMICS OF OCEAN POLICY IN THE ERA OF EXTENDED JURISDICTION*
Presiding GIULIO PONTECORVO, Columbia University
Papers JAMES CRUTCHFIELD, University of Washington
The Economics of U S Ocean Policy
PARZIVAL COPES, Simon Fraser University
The Economics of Marine Fisheries Management in the Era of Extended Jurisdiction The Canadian Perspective
VLADIMIR KACZYNSKI, Sea Fisheries Institute, Gdynia, Poland
The Economics of the Eastern Bloc's Ocean Policy
MAURICE WILKINSON, Columbia University
The Economic Analysis of Ocean Resource Problems

10 15 A.M RECENT DEVELOPMENTS IN THE ECONOMICS OF INFORMATION*
Presiding LESTER TELSER, University of Chicago
Papers JOHN G RILEY, University of California-Los Angeles
Alternative Signalling Equilibrium Concepts
STEPHEN A ROSS, Yale University
Equilibrium and Agency
CHARLES A WILSON, University of Wisconsin
Equilibrium and Adverse Selection
Discussants STEVEN C SALOP, Civil Aeronautics Board
ARTHUR RAVIV, Carnegie-Mellon University
STEVEN SHAVELL, Harvard University

10:15 A.M. NEW DEVELOPMENTS IN LABOR ECONOMICS AND THEIR IMPLICATIONS FOR POLICY ANALYSIS (Joint Session with the Industrial Relations Research Association)
Presiding: MELVIN W. REDER, University of Chicago
Papers JAMES HECKMAN, University of Chicago
Labor Supply
DANIEL HAMERMESH, Michigan State University
Multi-Labor Demand Functions and their Application to Policy Analysis
R. SMITH, Cornell University
Compensating Wage Differentials and Public Policy A Review

Discussants: RICHARD FREEMAN, Harvard
JOSEPH HOTZ, Carnegie-Mellon University
EDWARD LAZEAR, University of Chicago

10:15 A.M. HOUSING DEMAND (Joint Session with the Econometric Society)
Presiding: EDWIN MILLS, Princeton University
Papers: BRYAN ELLICKSON, University of California-Los Angeles
Hedonic Theory and Housing Markets
STEPHEN MAYO, Abt Associates
Theory and Estimation in the Economics of Housing Demand
JOHN PITKIN AND JEROME ROTHENBERG, Massachusetts Institute of Technology
Demand, Supply, and Market Interaction in a Segmented Urban Housing Market
Discussants: DENNIS CARLTON, University of Chicago
HENRY POLAKOWSKY, University of Washington
DAVID SEGAL, Harvard University

10 15 A.M. MACROECONOMICS: AN APPRAISAL OF THE NON-MARKET-CLEARING PARADIGM*
Presiding HERSCHEL I. GROSSMAN, Brown University
Papers ROBERT J. BARRO, University of Rochester
Keynesian Economics and Other Youthful Indiscretions
PETER HOWITT, University of Western Ontario
Short-Run Theory With or Without Market Clearing
HERSCHEL I GROSSMAN, Brown University
Why does Aggregate Employment Fluctuate?
Discussants JOSEPH OSTROY, University of California-Los Angeles
ROBERT M SOLOW, Massachusetts Institute of Technology

10.15 A.M ROUND TABLE ON THE ECONOMIC OUTLOOK (Joint Session with the American Finance Association)
Presiding. BURTON G. MALKIEL, Princeton University
Panelists. (To be announced)

10.15 A.M HEALTH MANPOWER RESEARCH (Joint Session with the Health Economics Research Organization)
Presiding LYNN E JENSEN, American Medical Association
Papers PAUL I FELDSTEIN, University of Michigan, AND CHARLES ROEHRIG, Policy Analysis, Inc
An Econometric Model of the Dental Sector
CHARLES R LINK AND RUSSELL F SETTLE, University of Delaware
The Supply of Married Professional Nurses
RICHARD L ERNST, University of Southern California, JOHN S GREENLEES, Bureau of Labor Statistics, AND DONALD F YETT, University of Southern California
U S and Foreign Medical School Graduates Comparison of Initial Career Choices
Discussants DONALD R HOUSE, American Dental Association
ROBERT T DEANE, Applied Management Sciences, Inc.
JAMES N HAUG, American College of Surgeons

2:00 P.M THE ACADEMIC LABOR MARKET FOR ECONOMISTS*
Presiding: C ELTON HINSHAW, Vanderbilt University
Papers CHARLES SCOTT, Old Marquette University
The Market for Ph.D. Economists. The Academic Sector
BARBARA REAGAN, Southern Methodist University
The Academic Labor Market for Women Economists
DAVID E AULT, GILBERT L. PUTNAM, Southern Illinois University-Edwardsville, AND THOMAS STEVENSON, St. Louis University
Efficiency in the Labor Market for Academic Economists
Discussants. HOWARD TUCKMAN, Florida State University
WILLIAM J. MOORE, University of Houston
MARION S. BEAUMONT, California State University-Long Beach

2:00 P.M. THE 1978 TAX BILL: A PANEL DISCUSSION
Presiding: JOSEPH A. PECHMAN, The Brookings Institution
Panelists: ALAN GREENSPAN, Townsend-Greenspan
WALTER W. HELLER, University of Minnesota
LAWRENCE R KLEIN, University of Pennsylvania
EMIL H. SUNLEY, JR., U.S. Treasury Department

2:00 P.M. ENERGY POLICY*
Presiding: ALAN S. MANNE, Stanford University
Papers. WALTER J. MEAD, University of California-Santa Barbara
(Title to be announced)
MICHAEL YOKELL, Solar Energy Research Institute
Federal Policy Toward New Sources of Energy
THOMAS LONG, University of Chicago, and LEE SCHIPPER, University of California-Berkeley
(Title to be announced)
Panelists GORDON COREY, Commonwealth Edison, Chicago
EDGAR FIEDLER, The Conference Board
WILLIAM HOGAN, Stanford University
DAVID STERNLIGHT, Atlantic Richfield Company

2:00 P.M. EQUITY THE INDIVIDUAL VS THE FAMILY (Joint Session with the Committee on the Status of Women in the Economic Profession)*
Presiding ANNE FRIEDLAENDER, Massachusetts Institute of Technology
Papers ROBERT A POLLAK, University of Pennsylvania, AND TERENCE WALES, University of British Columbia
Welfare Comparisons and Equivalence Scales
MARILYN E MANSER, State University of New York-Buffalo
Comparing Households with Different Structures The Problem of Equity
CAROL T F BENNETT, University of Texas-Austin
The Social Security Benefit Structure Equity Considerations of the Family as its Basis
Discussants ALICIA MUNNELL, Federal Reserve Bank of Boston
CLAIR VICKERY, University of California-Berkeley

2:00 P M. QUALITATIVE CHOICE (Joint Session with the Econometric Society)
Presiding RICHARD QUANDT, Princeton University
Papers DANIEL MCFADDEN, University of California-Berkeley
On the Use of Probabilistic Choice Models in Economics
JAMES HECKMAN, University of Chicago
Statistical Models for Discrete Panel Data
CHARLES MANSKI, Carnegie-Mellon University
Sample Design for Discrete Choice Analysis The State of the Art
Discussant DAVID WISE, Harvard University

2:00 P.M. TOPICS IN MACROECONOMICS
Presiding: ROGER WILLIAMS, St John's University
Papers DAVID CHARLES COLANDER, Vassar College
The Free Market Solution to Inflation
BRIAN A MARIS, Wright State University
Causality Between Money and Interest Rates
LAWRENCE S DAVIDSON AND JOHN M FINKELSTEIN, Indiana University
The Macroeconomic Impact of Variable Interest Rates
BENJAMIN WURZBURGER, Bank of Canada
A Neo-Keynesian Model of Nominal Wage Determination in Canada
Discussants. ROGER WILLIAMS, St John's University
JAMES E PRICE, Syracuse University
JAGDISH HANDA, McGill University

2.00 P.M. THE ECONOMICS OF VOLUNTARY STANDARDS
Presiding HAROLD MARSHALL, National Bureau of Standards
Papers. CAROL CHAPMAN, National Bureau of Standards
Bottle Safety. A Case Study of Industrial Self-Regulation
GARRETT VAUGHN, Senate Subcommittee on Antitrust and Monopoly
The Business of Supplying Information Does it Tend to Monopoly? The Case of Voluntary Industrial Standards
DAVID HEMENWAY, Harvard University School of Public Health
The Voluntary Standards Process and the Fire Problem
Discussants: PAUL GATONS, Consumer Product Safety Council
BARRY WEINGAST, Washington University
MALCOM GETZ, Vanderbilt University

2:00 P.M. MONETARY POLICY: ASSESSING THE BURNS YEARS (Joint Session with the American Finance Association)
Presiding. RICHARD C ASPINWALL, Chase Manhattan Bank, N.A.
Papers WILLIAM POOLE, Brown University
Burnsian Monetary Policy. Eight Years of Progress?
JAMES PIERCE, University of California-Berkeley
The Political Economy of Monetary Policy
Discussants: JERRY JORDAN, Pittsburgh National Bank
RAYMOND LOMBRA, Pennsylvania State University
THOMAS MAYER, University of California-Davis

8.00 P M PRESIDENTIAL ADDRESS
Presiding (To be announced)
Speaker TJALLING C KOOPMANS, Yale University
Economics Among the Sciences

9:30 P.M. BUSINESS MEETING

Thursday, August 31, 1978

8 00 A M ECONOMICS AMONG THE SCIENCES
Presiding ROBERT SOLOW, Massachusetts Institute of Technology
Panelists DANIEL MCFADDEN, University of California-Berkeley
JOHN DEUTCH, U S Department of Energy

8 00 A M ISSUES OF MONETARY POLICY*
Presiding WILLIAM POOLE, Brown University
Papers DALE HENDERSON, Federal Reserve Board
Monetary Policy and the Managed Float
BENNETT MCCALLUM, University of Virginia
The Current State of the Policy-Ineffectiveness Debate
JAMES PIERCE, University of California-Berkeley
A Case for Monetary Reform
Discussants JACOB FRENKEL, University of Chicago
WILLIAM BRAINARD, Yale University
FRANK MORRIS, Federal Reserve Bank of Boston

8 00 A.M ECONOMIC EDUCATION RESEARCH ISSUES AND ANSWERS*
Presiding KENNETH BOULDING, University of Colorado
Papers BURTON WEISBROD, University of Wisconsin
Research on Economic Education Is It Asking the Right Questions?
THOMAS JOHNSON, North Carolina State University
Research on Economic Education How Well is it Answering the Questions Asked?
Discussants PETER O STEINER, University of Michigan
ZVI GRILICHES, Harvard University

8 00 A.M. EVALUATING THE 1977 STIMULUS PACKAGE*
Presiding: DONALD NICHOLS, U S Department of Labor
Papers: MICHAEL L WACHTER, University of Pennsylvania
(Title to be announced)
EDWARD M. GRAMLICH, University of Michigan
(Title to be announced)

8.00 A M. ECONOMIC DEVELOPMENT TRADE ASPECTS*
Presiding ANNE O. KRUEGER, University of Minnesota
Papers RONALD FINDLAY, Columbia University
Economic Development and the Theory of International Trade
HOSSEIN ASKARI, WITH JOHN CUMMINGS AND GUNTER RICHTER, Graduate School of Business and Center for Middle Eastern Studies, University of Texas-Austin
The Efficiency of LDC Trading Patterns. The Iranian Experience
VITTORIO CORBO, Concordia University and Institute of Applied Economic Research
Trade and Employment: Chile in the 1960's

Discussants: ROBERT BALDWIN, University of Wisconsin-Madison
HENRY BRUTON, Williams College
JEFFREY NUGENT, University of Southern California

8:00 A.M. FACTORS IN RESIDENTIAL LOCATION
Presiding: PHILLIP WEITZMAN, National Social Science and Law Project, Inc.
Papers: HIRSCHEL KASPER, Oberlin College and Cornell University
Toward Estimating the Incidence of Journey to Work Costs
BARRY MCCORMICK, University of Cambridge
Rent Control, Labor Mobility, and Unemployment
JANICE FANNING MADDEN, University of Pennsylvania
The Effect of Residential Location, Job Location, and the Scheduling of Work Shifts on Labor Supply
Discussant: PHILLIP WEITZMAN, National Social Science and Law Project, Inc.

8:00 A.M. MACROECONOMIC MARTINGALES (Joint Session with the Econometric Society)
Presiding: DANIEL O'NEILL, Georgia Institute of Technology
Papers: DANIEL O'NEILL, Georgia Institute of Technology
Sources of Macroeconomic Martingales and Tests of the Martingale Property
FREDERIC S. MISKHIN, University of Chicago
Efficient Markets Theory: Implications for Monetary Theory
ROBERT J. SHILLER, University of Pennsylvania
The Volatility of Long-Term Interest Rates and Efficient Markets Theory
Discussants: MICHAEL HAMBURGER, Federal Reserve Bank of New York
GARY SKOOG, University of Chicago
CHRISTOPHER SIMS, University of Minnesota

8:00 A.M. ASPECTS OF INTERNATIONAL ECONOMICS
Presiding: JAMES M. HOLMES, Arizona State University
Papers: RICHARD BERNER ET AL., Federal Reserve System
Report on a Multicountry Model
C. JEVONS LEE, Wesleyan University
Export and Import Functions of Small Economies: A General Equilibrium Approach
JOSEPH A. HASSON, U.S. Department of Energy
Macroeconomic Effects of the Optimal Tariff
JEFFREY SACHS, Harvard University
Wage Indexation, Flexible Exchange Rates, and Macroeconomic Policy
CARL VAN DUYNE, Williams College
The Macroeconomic Effects of Commodity Market Disruptions in Open Economies

8:00 A.M. "RENT SEEKING" (Joint Session with the Public Choice Society)
Presiding: ROBERT TOLLISON, Virginia Polytechnic Institute and State University
Papers: GORDON TULLOCK, Virginia Polytechnic Institute and State University
The Backward State
MANCUR OLSON, University of Maryland
Rent Seeking and Growth
TERRY L. ANDERSON AND P. J. HILL, Montana State University
Rent Seeking in Nineteenth Century America
Discussants: GEOFFREY BRENNAN, Virginia Polytechnic Institute and State University
RICHARD B. MCKENZIE, Clemson University
ROBERT E. MCCORMICK, Graduate School of Management, University of Rochester

10:15 A.M. ENTREPRENEURSHIP
Presiding: J. FRED WESTON, University of California-Los Angeles
Papers: GERALD O'DRISCOLL, Iowa State University
Rational Expectations and Entrepreneurship
MARIO RIZZO, New York University
Radical Uncertainty: The Source of True Economic Profit
ISRAEL M. KIRZNER, New York University
Alertness, Luck, and Entrepreneurial Profit
Discussants: T. W. SCHULTZ, University of Chicago
J. HUSTON MCCULLOCH, Boston College

10.15 A.M. MARKETING AND TRADE STRATEGIES FOR THIRD WORLD COUNTRIES: COMPETITION AND COOPERATION (Joint Session with the National Economics Association)

Presiding. ALEX O. WILLIAMS, University of Virginia

Papers STEPHANIE Y. WILSON, Abt Associates

The Impact of a Multinational Service Agency on Industry Structure in Less Developed Countries

GAVIN MICHAEL CHEN, U.S. Department of Commerce-OMBE, AND GERALD F. WHITTAKER, Chicago Economic Development Corporation

Nationalism and Economic Development Regional Cooperation and Economic Adjustments in the Caribbean

ERNEST L MURPHY, Howard University

The Impact of Agricultural Policies on Trade and Marketing of Coffee in Haiti

Discussants RONALD MULLER, American University

THOMAS D BOSTON, Atlanta University

DIANN PAINTER, Wellesley College

10 15 A.M REGIONAL COMPETITIVENESS

Presiding JOSEPH QUINN, Boston College

Papers JANET SPRATLIN YOUNG, Federal Reserve Bank of New York

Does it Pay to Work in New York? An Examination of Migration of Workers in the Financial Industry

JOHN J SIEGFRIED, Vanderbilt University

Minimizing AEA Convention Costs

PEARL KAMER, Nassau-Suffolk Regional Planning Board

The Competitiveness of U S. Urban Areas for Manufacturing: Sunbelt vs Nonsunbelt SMSAs

Discussants THOMAS J. KNIESNER, University of North Carolina

JOHN L BUNGUM, University of Wisconsin

JOSEPH QUINN, Boston College

10 15 A M LAW AND ECONOMICS (Joint Session with the Law and Society Association)

Presiding WARREN J. SAMUELS, Michigan State University

Papers DUNCAN KENNEDY, Harvard Law School

Two Phases of the Fetishism of Commodities

DAVID TRUBEK, Law School, University of Wisconsin

Government Regulation of Business in Capitalist Economies Neoclassical, Weberian, and Marxist Approaches

Discussants KARL DE SCHWEINITZ, Northwestern University

LAWRENCE S BACOW, Massachusetts Institute of Technology

10 15 A.M INHERITANCE OF HUMAN AND NONHUMAN WEALTH AND INCOME DISTRIBUTION

Presiding: (To be announced)

Papers GIAN S SAHOTA, Vanderbilt University, AND CARLOS A ROCCA, University of Sao Paulo

A More Comprehensive Theory of Personal Income Distribution, a Synthesis of Becker and Meade

PAUL L MENCHIK, University of Wisconsin

Unequal Material Inheritance and Intergenerational Mobility and Intragenerational Equality

JOSEPH E STIGLITZ, Oxford University

Inheritance, Taxation, and Wealth Inequality

JAMES ADAMS, Iowa State University

Human and Material Inheritance Within and Between Families

Discussants GARY S BECKER, University of Chicago

HAROLD WATTS, Columbia University

FINIS WELCH, University of California-Los Angeles and Rand Corporation

PAUL TAUBMAN, University of Pennsylvania

10 15 A M. EFFECTS OF INFLATION UNCERTAINTY

Presiding: ROBERT LUCAS, University of Chicago

Papers: RICHARD BOOKSTABER, Boston University

The Use of Authority Induced Uncertainty as a Tool of Monetary Policy

ALEX CUKIERMAN AND PAUL WACHTEL, New York University

Differential Inflationary Expectations

MAURICE D. LEVI, University of British Columbia, AND JOHN H MAKIN, University of Washington
Expectational Processes and Nonneutrality of Monetary Disturbances
Discussants. WILLIAM POOLE, Brown University
OLIVIER BLANCHARD, Harvard University

10.15 A.M. REGULATION OF NURSING HOMES (Joint Session with Health Economics Research Organization)
Presiding. LINDA A. SIEGENTHALER, National Center for Health Services Research
KENNETH L. HAMILTON, Georgia Institute of Technology
Mandatory Medicare Participation for Medicaid Certified Nursing Homes Cost-Benefit Policy Analysis
CHRISTINE BISHOP AND HOWARD BURNBAUM, Abt Associates
Government Regulation of Nursing Homes. The Case of Prospective Reimbursement
PATRICIA REAGAN, Massachusetts Institute of Technology
Regulating The Nursing Home Industry Some Unanswered Questions
Discussants DOUGLAS E. SKINNER, Applied Management Sciences, Inc.
JOHN S. GREENLEES, U.S. Bureau of Labor Statistics
PAUL B. GINSBURG, Duke University

10:15 A.M QUESTIONS OF DISCRIMINATION
Presiding ROBERT A MOFFITT, Mathematic Policy Research, Inc
Papers JAMES F RAGAN, Kansas State University, AND SHARON P. SMITH, Federal Reserve Bank of New York
Are Sex Differences in Layoff Rates Accompanied by Compensating Pay Differentials?
MARK R. KILLINGSWORTH, Barnard College
Racial and Institutional Differentials in Faculty Salaries in Public Institutions of Higher Education in Tennessee
MASANORI HASHIMOTO AND LEVIS A. KOCHIN, University of Washington
A Bias in Estimating Racial Differences
Discussants ROBERT A. MOFFITT, Mathematica Policy Research, Inc.
PETER GOTTSCHALK, Bowdoin College

10.15 A.M. THE CHANNELS OF INFLUENCE OF TOBIN-BRAINARD'S "Q" ON INVESTMENT (Joint Session with the American Finance Association)
Presiding. MICHAEL A. KLEIN, Indiana University
Papers. JOHN CICCOLO, Boston College, AND GARY FROMM, SRI International
"*q*" and the Theory of Investment
BURTON G. MALKIEL, Princeton University, GEORGE M VON FURSTENBERG AND HARRY S. WATSON, Indiana University
Industry "*q*'s", Corporate Investment, and Wage and Price Changes
Discussants MICHAEL C LOVELL, Wesleyan University
RUSSELL SHELDON, U.S Department of Commerce

Editor's Note:
*Papers from sessions marked with an asterisk will be published in the Papers and Proceedings issue of this *Review*.

ERA and the Site of Annual Meetings

A LETTER TO THE MEMBERS OF THE AMERICAN ECONOMIC ASSOCIATION

From Tjalling C. Koopmans, President

Over four hundred persons have written (or signed) letters to me about the plans of the American Economic Association to hold its meetings in 1978 and 1979 in states that have not ratified the Equal Rights Amendment (ERA) to the U.S. Constitution. A large majority of them objected to these plans, although some urged disregarding ERA in site selection. Upon consultation with President-elect Robert Solow, I have thought that the intense interest in the subject expressed in the correspondence warranted reconsideration by the Executive Committee at its meeting on March 17, 1978.

Unable to write individual replies to all the communications I have received, I wish to give a report on the way the Executive Committee reached its decision of March 17 and the reasons behind it.

The original decisions to meet in Chicago in August 1978 and Atlanta in December 1979 were made in 1971 and 1973 when even in retrospect it is hard to see how ERA could have been taken into account. According to the most recent report of the Secretary, the schedule for subsequent meetings is September 1980 in Denver, and December 1981 in Washington, D.C. For 1982 the site will be New York. Tentative plans are to hold the 1983 meetings in San Francisco.

The question of changing the sites for 1978 and/or 1979 was first discussed in the Executive Committee at its meeting on December 27, 1977. According to the minutes of that meeting,

> On behalf of the Committee on the Status of Women in the Economics Profession and other concerned members of the Association, Friedlaender proposed that the Association change the sites for the 1978 and 1979 meetings because they are currently scheduled for States (Illinois and Georgia) which have not ratified the Equal Rights Amendment to the U.S. Constitution. Counsels Turner and Raskind advised that such an action would conflict with Article 3 of the Association's Certificate of Incorporation—"The Association as such will take no partisan attitude, . . ." After a consideration of the possible economic and legal consequences of moving the sites, it was moved that immediate inquiries be made to determine the feasibility and desirability of withdrawing from Atlanta in 1979. The motion failed. It was understood that this action did not imply any position on the Equal Rights Amendment.

The vote not to take action evoked the communications which led to the further consideration on March 17. The following summary of the discussion includes almost verbatim the paragraphs from the draft minutes for this agenda item as circulated for approval to the Executive Committee members by the Secretary. Ann Friedlaender and Susan Rose-Ackerman, attending the meeting as guests, proposed that the Executive Committee (1) explore the possibility of moving the Atlanta meetings to a state that has ratified ERA and (2) issue a statement to the effect that after 1979 the Association will not meet in states that have not ratified the ERA so long as passage of the Amendment is a live issue.

After the guests made their presentation and responded to questions, the Executive Committee discussed the subject at length. I distributed a tabular summary of the messages received through March 15. A further condensation of this summary is appended. I also append a straddling sample of four of

SUMMARY OF LETTERS RECEIVED THROUGH MARCH 15, 1978

Number of Signatures by Type of Letter and Sex of Signatory	Oppose Meeting in Chicago or Atlanta	Only Want No Meetings in Nonratifying States Thereafter	MIT[a]	Want to Ignore ERA in Site Selection	Oppose ERA	Want Poll of Members	Will Not Come to Chicago or Atlanta
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Individual							
57 Women	40	9		7	3	9	1
28 Men	24	-		3	1	2	1
3 Unidentified	1	1		1	-	1	-
88 Total	65	10		11	4	12	2
Collective (26 Letters)							
94 Women	89	3	1	1		16	
224 Men	176	11	22	15		39	
25 Unidentified	21	-	4	-		5	
343 Total	286	14	27	16		60	
All Letters							
151 Women	129	12	1	8	3	25	1
252 Men	200	11	22	18	1	41	1
28 Unidentified	22	1	4	1	-	6	-
431 Total	351	24	27	27	4	72	2

Note: (1) = (2) + (3) + (4) + (5) + two individual letters under (7).
[a]Position defined in attached collective letter.

the letters, two individual and two collective ones, selected for possessing in varying degrees the merits of brevity, articulateness, moderation, and effective advocacy of the position supported. The two collective letters were read aloud at the meeting. All letters were available for the members of the Executive Committee to inspect for themselves.

As far as could be told from the ensuing discussion, there was full agreement on two points: it was too late to change the site of the meetings scheduled for August 29–31, 1978, in Chicago; and the provision in the charter forbidding the Association from taking a position on partisan issues was wise.

Those members favoring a move out of Atlanta and a policy of refusing to meet in states that have not ratified ERA expressed the view that the commitment to meet in Atlanta may not be a firm contractual obligation, that the likelihood of the Atlanta hotels bringing a lawsuit under the laws of contract or the antitrust laws was minimal, that the Association could take into account the preferences of its members as to meeting sites without violating its charter (i.e., without taking a position on ERA itself), that meeting in Atlanta might discourage attendance by young women economists most in need of the benefits of attending or put them in the awkward position of compromising their principles, that meeting in Atlanta would be inconsistent with the position taken by the Association in favor of equal treatment of women in the economics profession, that ERA is not controversial within the Association and has been ratified by states with 70 percent of the population of the United States, that the Association can most effectively make known the position of its members on ERA by refusing to meet in states that have not ratified it, and that in view of the number of associations that have taken stands of the kind being proposed for the AEA, failure to act in this

way would be construed as a negative position on ERA itself.

Opposition to the proposals centered on three arguments: that it is illegitimate for the Association to use economic pressure to bring about political change, that adoption of the proposals would be taken as endorsement by the AEA of ERA, i.e., would constitute taking a position on a political issue, and that selecting sites to further a political measure would set a dangerous precedent. It appeared from the discussion, however, that there was considerable agreement that the Association in future selection of meeting sites should take account of the preferences of its members, regardless of how they have been arrived at.

After thorough discussion, the Executive Committee by a decisive majority voted to take no action regarding Chicago or Atlanta but to consider how and to what extent members' preferences concerning site selection could be taken into account in the future.

In the chair I limited my participation in the discussion to matters of procedure and clarification, and to confining the flow of discussion to one speaker at a time. I also did not vote, holding myself in reserve for the event of a tied vote. Having been spared that agony by a substantial margin, allow me here to express my concern about a possible self-imposed obligation of the AEA to take account of all "preferences of members regardless of how arrived at." The danger of proliferation of such issues is described with verve in the letter of Mary Alice Shulman. Briefly, I am concerned about the possible use of the Association as an instrument for the achievement of goals that, however desirable they may be, go far beyond those to which the Association is dedicated. Such a development is likely to diminish the effectiveness of the Association with regard to its own distinctive goals.

TJALLING C. KOOPMANS, *President*
April 10, 1978

A Sample of Two Individual and Two Collective Letters

1. This letter is in regard to AEA choices of annual meeting sites and times. While the AEA should strive to be apolitical, by its very nature it should also be exemplar in its posture relative to certain basic values. If efficiency to economists is a prime good, then the underlying basis must surely be the normative position that attempts to improve the potential lot of people on this earth are worthwhile. Having annual meetings in the states of Illinois and Georgia, it seems to me, has as much political overtones as not holding meetings in these places. In either case, an example is set. And, in any case, the views of AEA rank and file membership should be reflected through known consent to both location sites and seasons. I request that the development of a process for this reflection get high priority in association future business.

In the meantime, I appreciate the somewhat landlocked situation of the 1978 AEA annual meeting in Chicago, although I think it is deplorable. But the 1979 Atlanta meeting surely demands rethinking. AEA patronage of states which patently ignore the vital necessity of continued vigilance in the apparently never ending march to secure human rights, seems unnecessary, as well as inappropriate.

Signed: MIETZL MILLER, Department of Economics, Lamar University, February 8, 1978

2. I do not agree with CSWEP that the American Economic Association should cancel conventions in states which have not ratified ERA. I strongly support ERA but believe this kind of boycott action accomplishes little and injures many innocent third parties, some of whom may themselves be ERA supporters.

Additionally, this proposed action would set a dangerous precedent to locate future conventions on the basis of whether the state had a clean bill of health on political

issues of interest to some portion of our membership migrant worker rights, abortion laws, labor laws, civil liberty issues, etc., etc., etc. I urge you to reject this misguided proposal.

Signed: MARY ALICE SHULMAN, Department of Economics, Northwestern University, February 12, 1978

3. I suppose because of my position as chairman, I was sent a copy of the CSWEP letter urging that the AEA, at least in the future, discriminate in its choice of meeting sites against cities that happen to be located in states that have not ratified the Equal Rights Amendment.

The letter contains no suggestion that our members will suffer indignities, or that they will not be accorded the same rights and freedoms that they have in other states. There is thus no parallel to the situation that used to prevail with respect to certain cities in the South, which could properly (at the time) be boycotted in defense of the personal dignity of black and other dark skinned (e.g., Indian) members of the Association.

The only reason to consider a state's stance on ERA as having any bearing on the location of the AEA meetings is political. I believe that the Executive Committee acted properly in refusing to "explore cancelling the Chicago or Atlanta meetings or to make a formal statement that future conventions will not be held in states that have not ratified ERA" on the ground that political action is prohibited by the Association's Charter. I hope that the Association will continue in the future to follow the wise example of the Executive Committee.

Signed: ARNOLD C. HARBERGER and eight cosigners, Department of Economics, University of Chicago, February 10, 1978

4. We are distressed that the American Economic Association is holding its next two meetings, respectively, in Illinois and Georgia, neither of which has ratified the Equal Rights Amendment. While we realize that it is probably too late to change the Chicago meeting, we think that the Association should explore the possibility of changing the Atlanta meeting. In addition, we think that the Association should make a formal statement that it will not hold future conventions in states that have not ratified ERA. While we recognize that the Association is constrained by its by-laws and contractual obligations, we understand that many other associations (most notably the AAAS) have managed to change their meeting sites, and we urge the American Economic Association to take a positive stance on this important issue.

Signed: PAUL A. SAMUELSON and twenty-six cosigners, Department of Economics, Massachusetts Institute of Technology, March 9, 1978

ANNOUNCEMENTS

The ninety-first annual meeting of the American Economic Association will be held in Chicago, IL, August 29–31, 1978. Please note that there will not be an Employment Center at the August meetings.

The 1978 Employment Center will be held December 28–30 at the Conrad Hilton in Chicago, Illinois Operating hours will be December 28, 10:00 A.M.–5:00 P.M., December 29th and 30th, 9:00 A.M.–5:00 P.M. Hotel reservation cards will be mailed to you upon receipt of your placement form.

Requests for placement forms should be made to Ms. Kathy Nichols, National Registry for Economists, Illinois State Employment Service, 40 West Adams Street, Chicago, IL 60603 between September 1 and November 15. *Note* No forms will be mailed prior to September; however they will be available at the American Economic Association booth at the annual meeting. Completed forms *must* be returned by December 1. You do not have to attend the meeting to have your listing on file. There is no registration fee.

Resolutions for Consideration at the Annual Business Meeting

The Executive Committee at its meeting on March 8, 1974, voted to require that, to be considered at the annual business meeting, proposed resolutions must be submitted to the Secretary at least one month in advance in writing with signatures of the proposer and the second, both of whom must be members in good standing The Secretary will reproduce the proposed resolutions and make copies available in advance of the meeting The next business meeting will begin at 9.30 P.M. on August 30, 1978 The deadline for proposed resolutions is accordingly July 30 They should be sent to the Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

Nominations for AEA Officers

The Electoral College on March 17 chose Moses Abramovitz as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1978. Other nominees (chosen by the nominating committee) are: for Vice-President (two to be elected), Hollis B. Chenery, Arnold C. Harberger, Jack Hirshleifer, and Irma Adelman; for members of the Executive Committee (two to be elected), Henry J. Aaron, Samuel Bowles, Zvi Griliches, and Daniel McFadden.

Under a change in the bylaws as described in the *Papers and Proceedings* of this *Review*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

1979 Nominating Committee of the AEA

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Robert Solow has appointed a Nominating Committee for 1979 consisting of Franco Modigliani, Chairman; Andrew Brimmer, Rachel Chamberlain, George Daly, Hirschel Kasper, John Riley, and Sherwin Rosen. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signature and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D degree or its equivalent, and must be citizens or permanent residents of the United States To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are meetings scheduled between July and October, March 1 for meetings scheduled between November and February, July 1 for meetings scheduled between March and June, November 1 Please request application forms by writing directly to the ACLS (Attention Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting Even when plans are

incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

There will be a session at the 1978 Economic History Association meetings in Toronto, Ontario, September 14 16, devoted to reports on dissertation research by six to eight students who will receive their Ph D in economic history by the end of the summer 1978 The session will be handled by a committee consisting of Professors Michelle McAlpin, Mary Yaeger, and Claudia Goldin, Chairperson Individuals wishing to be considered for participation should send two copies of a 2,000 word abstract of their dissertation by July 15, 1978 to Professor Claudia Goldin, Department of Economics, Princeton University, Princeton, NJ 08540 The names of those selected will be announced by August 1

The North American Economic Studies Association will meet December 28 29, 1978, in Mexico City. The meeting will be hosted by La Universidad de las Americas, A C Papers on all aspects of the economics of the nations of North America including the Caribbean Islands will be considered for inclusion on the program. Papers in either English or Spanish are welcome Economists who wish to present papers or to serve as session chairpersons or discussants may contact Dr. Joseph Horton, Department of Economics and Business, Slippery Rock State College, Slippery Rock, PA 16057, or Dr J Hodgson, Centro de Estudios Universitarios, Universidad de las Americas, A.C. Hamburgo 250, Mexico 6 D.F Mexico To be considered for inclusion on the program, abstracts of papers must be received by July 31, 1978

The jointly sponsored Ford and Rockefeller Foundations' Research Program is interested in receiving proposals focusing on population policy as it relates to social and economic development Of particular interest to this year's program are proposals that may help in closing the gap between research and policy planning on development issues Submissions are encouraged on a broad range of topics. Research areas are, for example: (1) interrelation of population policies and other development policies, particularly with respect to food, energy, and employment; (2) impact of government programs in such areas as rural development, health, education, housing, social security, and transportation on rural to urban migration and/or fertility; (3) interrelations among infant and child mortality, nutrition, age at marriage, socioeconomic factors, and fertility; (4) simulation of the economic, social, and/or demographic consequences of alternative population and development policies at various levels of assumed effectiveness; (5) policy implications of internal and international migration trends. The deadline for submission of proposals is July 1, 1978, and awards will be announced in December. The proposed research may begin on or after Jan. 1, 1979. For further program information, write to The Ford and Rockefeller Foundations' Research Program on Population and Development Policy, The Ford Foundation, 320 East 43rd Street, New York, NY 10017.

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics announce the 1978 Leonard J Savage Award of $500 for an outstanding doctoral dissertation in the area of Bayesian econometrics and statistics. To be considered, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1978, accompanied by a short letter from the supervisor summarizing the main results of the dissertation Dissertations completed after January 1, 1975 are eligible to considered. An evaluation committee will be appointed by the board of the Leonard J Savage Memorial Trust Fund (S E Fienberg, S Geisser, J B. Kadane, E. E Leamer, J W. Pratt, and A Zellner, chairman) to evaluate submitted dissertations. Contact Professor Arnold Zellner, Graduate School of Business, University of Chicago, 5836 S Greenwood Avenue, Chicago, IL 60637

The winner of the 1977 Savage Award is Charles Holt, University of Minnesota, whose doctoral dissertation, "Bidding for Contracts," was completed at Carnegie-Mellon University. Honorable mention was given to Robert Shore, "A Bayesian Approach to the Spectral Analysis of Stationary Time Series," completed at Carnegie-Mellon University. The 1977 evaluation committee was R E Kihlstrom, chairman, C Barry, E E Leamer, R A. Olshen, and J W Pratt.

The British Politics Group (*BPG*) intends to issue the third edition of its Register of Current Research in early 1979 The research Register lists ongoing work on any aspect of British politics Scholars who are not members of *BPG* but who wish to be included in the Register should contact William D. Muller, Editor, *BPG* Research Register, c/o Political Science Department, State University of New York, Fredonia, NY 14063. Copies of the Register will be sent to members of the *BPG* in early 1979. Nonmembers may purchase a copy from Jorgen Rasmussen, Executive Secretary *BPG*, c/o Political Science Department, Iowa State University, Ames, Iowa 50011.

The directors of the Harold Innis Foundation are establishing a clearing house for studies relating to Harold Innis and his work. To this end, the directors are soliciting names of scholars interested in the work of Innis, and copies of or references to any material, completed or in progress, related to Innis' work They are concerned with his interest in Canadian economic

history, political economy, and the history of communications. The clearing house will in future provide a central resource for scholars working in all of these spheres. Address suggestions to the Innis Foundation, c/o Innis College, University of Toronto, Toronto, Ontario, M5S 1J5

The *Journal of Post Keynesian Economics* is committed to the principle that the cumulative development of economic theory is possible only when the theory is continuously subject to challenge, both in terms of its ability to explain the real world and to provide a reliable guide to public policy Post Keynesian economics is to be broadly interpreted, spotlighting new problems and revealing new theoretical prospectives, this view is consonant with Keynes' vision of the open-ended nature of economic study An objective of the *JPKE* is the encouragement of publication by younger members of the profession For information on submission of manuscripts, write to Professor Paul Davidson, Rutgers University, Winants Hall, Rm 105, New Brunswick, NJ 08903

The Conference Group on Nordic Society is continuing its efforts to locate all individuals interested in social science research and teaching on the Scandinavian countries A second wave of questionaires is currently being circulated to individuals actively engaged in research and teaching The questionaire is designed to construct a revised directory of published and current research If you teach or do social science research on Scandinavia and have not received a copy, please write for a questionaire Write if you are only interested in securing the current Directory, or finding out about visiting lecturers, research opportunities, and workshops Address all inquiries to Professor Robert B Kvavik, University of Minnesota, Department of Polical Science, 267 Nineteenth Avenue, South, 1414 Social Sciences Bldg, Minneapolis, MN 55455

Saitama University announces the opening of a visiting foreign professor chair at its Institute for Policy Science during the academic year 1978 79 Those interested in applying should write to Mizuho Ogawa, Director, Institute for Policy Science, Saitama University, Urawa, Saitama 338, Japan This announcement was supplied by the Council for International Exchange of Scholars, Suite 300, Eleven Dupont Circle, Washington, D.C 20036 (202+833-4950).

The Sixth Annual History of Economics Society Conference will be held at the University of Illinois, May 24-26, 1979 All communications should be sent to Professor Royall Brandis, Department of Economics, University of Illinois, Urbana, IL 61801 The purpose of the History of Economics Society is to promote interest and inquiry into the history of economics and related parts of intellectual history. Information about membership in the Society, which may be combined with a reduced-price subscription to the journal *History of Political Economy*, can be obtained from James L Cochrane, Secretary-Treasurer, History of Economics Society, Department of Economics, University of South Carolina, Columbia, SC 29208.

Fellowship Announcement

The Latin American Program of the Woodrow Wilson International Center for Scholars will award about five postdoctoral fellowships in 1979 for research by social scientists and humanists on Latin America, the Caribbean, and inter-American affairs. Interest centers on a number of central themes: the interplay between the international economic order and domestic political and economic choices in Latin America and the Caribbean; the nature and evolution of U.S.-Latin American relations, and Latin America's international role, the causes and dynamics of authoritarianism in Latin America; the interplay between cultural traditions and political institutions in the region, the history of ideas in Latin America as they bear on contemporary public policy choices, and the dynamics and viability of alternative development models in Latin America and the Caribbean.

The program is residential and fellows are expected to devote full time at the Center to the major research project proposed in their applications Appointments normally extend from four months to a year in duration. The competition is open and international, with applications welcomed from any country The deadline for the 1979 competition is October 1, 1978 For further information and application forms, write Alexander Wilde, Research Associate, Latin American Program, Woodrow Wilson International Center for Scholars, Smithsonian Institution Bldg, Washington, D.C 20560.

The National Association of Affiliated Economic Education Directors will meet in Portland, Oregon, October 5 8, at the Benson Hotel. A session on research papers in economic education is scheduled for October 6. Anyone interested in economic education research is invited to submit an abstract (with or without accompanying manuscript) for review. Those interested in serving as discussants are encouraged to submit their names, addresses, and professional affiliations Please submit abstracts and other materials by July 15, 1978, to Robert J. Highsmith, Council on Economic Education in Maryland, Towson State University, Towson, MD 21204.

The 1978 edition of the Binational Science Foundation (BSF) brochure, "Applications for Grants and Guidelines for Recipients," is available. It describes the BSF support program for U S Israel cooperative scientific research projects covering a wide spec-

trum of scientific disciplines. In the social sciences, emphasis will be given to economic theory, monetary and fiscal economics, labor economics. It also describes the Professor E. D. Bergmann Memorial Research Grants to young scientists. The BSF normally finances cooperative research performed substantially in Israel. Awards are made in Israeli currency. Brochures, application forms, and answers to specific questions may be obtained from Dr R. Ronkin, Division of International Programs (U.S. Israel Binational Science Foundation), Washington, D C 20550.

Death

Robert Aaron Gordon, professor emeritus, University of California-Berkeley, Apr 7, 1978

Retirements

Bertis E Capehart, director, education department, American Iron and Steel Institute, Nov 1, 1977

John D Gaffey, economist, Antitrust Division, U S Department of Justice, Los Angeles Field Office, Feb 1977

George R Lent, senior advisor, fiscal affairs department, International Monetary Fund, July 1977

Visiting Foreign Scholars

D Johannes Jüttner, Macquarie University, Australia visiting professor, West Virginia University, Aug 1977

Promotions

Richard X Chase professor of economics, University of Vermont, Sept 1, 1977.

Michael F d'Amico associate professor of marketing, University of Akron, Sept. 15, 1977

J Eric Fredland associate professor, department of economics, U S Naval Academy, Sept 1977

Robert F. Gemmill associate director, Division of International Finance, Board of Governors of the Federal Reserve System, June 1977.

John M. Godfrey research officer, Federal Reserve Bank of Atlanta, Jan 1, 1978

J. Kevin Green associate professor of economics, University of the South, Sept. 1, 1977

Josef Gruber. professor of statistics and econometrics, department of economics, Fernuniversität Hagen, West Germany, Jan 1977

George B. Henry associate director, Division of International Finance, Board of Governors of the Federal Reserve System, June 1977

Sungwoo Kim professor of economics, Northeastern University, Sept. 1977.

Roger D. Little: associate professor, department of economics, U.S. Naval Academy, Feb. 1977.

Francis P. Mulvey: assistant professor of economics, Northeastern University, Sept. 1977.

Robert P. Parker. chief, National Income and Wealth Division, Bureau of Economic Analysis, U.S. Department of Commerce, Nov 23, 1977.

John E Reynolds counselor, Division of International Finance, Board of Governors of the Federal Reserve System, June 1977

P K Sawhney associate professor of economics, Northeastern University, Sept 1977

Charles J Siegman associate director, Division of International Finance, Board of Governors of the Federal Reserve System, June 1977

Edwin M Truman director, Division of International Finance, Board of Governors of the Federal Reserve System, June 1977

Administrative Appointments

Herman A Berliner associate provost for budget and curriculum, Hofstra University, Sept 1, 1977.

L A D Dellin chairman, department of economics, University of Vermont, Sept 1, 1977.

Dimitri B Papadimitriou vice president of finance and management, Bard College, Dec 1, 1977

James G Ward director of economic research, American Federation of Teachers, Oct 1, 1977

Appointments

Gerald E Auten assistant professor of economics, Bowling Green State University, Sept 1977

Malcolm D Bale economist, World Bank, Jan 1978

Christopher F Baum, University of Michigan assistant professor of economics, Boston College, Sept 1977

Matthew D Berman, Yale University assistant professor, LBJ School of Public Affairs, University of Texas-Austin, Sept 1977

Leigh B Boske, National Transportation Policy Study Commission assistant professor, LBJ School of Public Affairs, University of Texas-Austin, Sept 1977

Michael J Boskin research associate and program director, social insurance, National Bureau of Economic Research, Sept 1977

David F Bradford research associate and program director, business taxation and finance, National Bureau of Economic Research, Sept 1977

Gerard Caprio economist, Division of International Finance, Board of Governors of the Federal Reserve System, Oct 1977

Douglas C Coate. research associate, National Bureau of Economic Research, Sept 1977.

Thomas A Connors economist, Division of International Finance, Board of Governors of the Federal Reserve System, Sept 1977

Rudiger Dornbusch research associate, National Bureau of Economic Research, Sept 1977

Robert W. Fogel research associate and program director, growth of American economy, National Bureau of Economic Research, Sept. 1977.

Richard B. Freeman. research associate and program director, labor studies, National Bureau of

Economic Research, Sept. 1977.

Richard T. Freeman: economist, Division of International Finance, Board of Governors of the Federal Reserve System, Sept. 1977.

Benjamin M. Friedman: research associate, National Bureau of Economic Research, Sept 1977

Guy C. Grenier. agricultural economist, U S Department of Agriculture, Nov 1976

Robert E. Hall research associate and program director, economic fluctuation, National Bureau of Economic Research, Sept. 1977

F. Reed Johnson: assistant professor, department of economics, U.S. Naval Academy, Sept 1977

Jacob J. Kaufman, Pennsylvania State University: metropolitan director and professor, New York State School of Industrial and Labor Relations, Cornell University, July 1, 1977

James L Medoff research associate, National Bureau of Economic Research, Sept 1977

William G. Nelson. director of implementation services, Office of Research, University of Pittsburgh.

Sherwin Rosen. research associate, National Bureau of Economic Research, Sept. 1977.

Daniel J. Sullivan assistant professor of economics, University of Vermont, Sept 1, 1977

Robert D Weaver: assistant professor of agricultural economics, Pennsylvania State University, July 1, 1977

John Scott Winningham: financial economist, Federal Reserve Bank of Kansas City, Jan. 4, 1978

Leaves for Special Appointment

George L Miller, Northern Illinois University: AACSB fellow, economist, Office of Macroeconomic Impact, Federal Energy Administration, Department of Energy, July 1, 1976-July 1, 1978.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style

A Please use the following categories

1- Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4 Promotions
5 Administrative Appointments
6 New Appointments
7 Leaves for Special Appointments (NOT Sabbaticals)
8 Resignations
9- Miscellaneous

B Please give the name of the individual (SMITH, Jane W), her present place of employment or enrollment, her new title (if any), and the date at which the change will occur.

C Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows. *March*, November 1, *June*, February 1, *September*, May 1, *December*, August 1

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912

HARRY G. JOHNSON

DISTINGUISHED FELLOW

1977

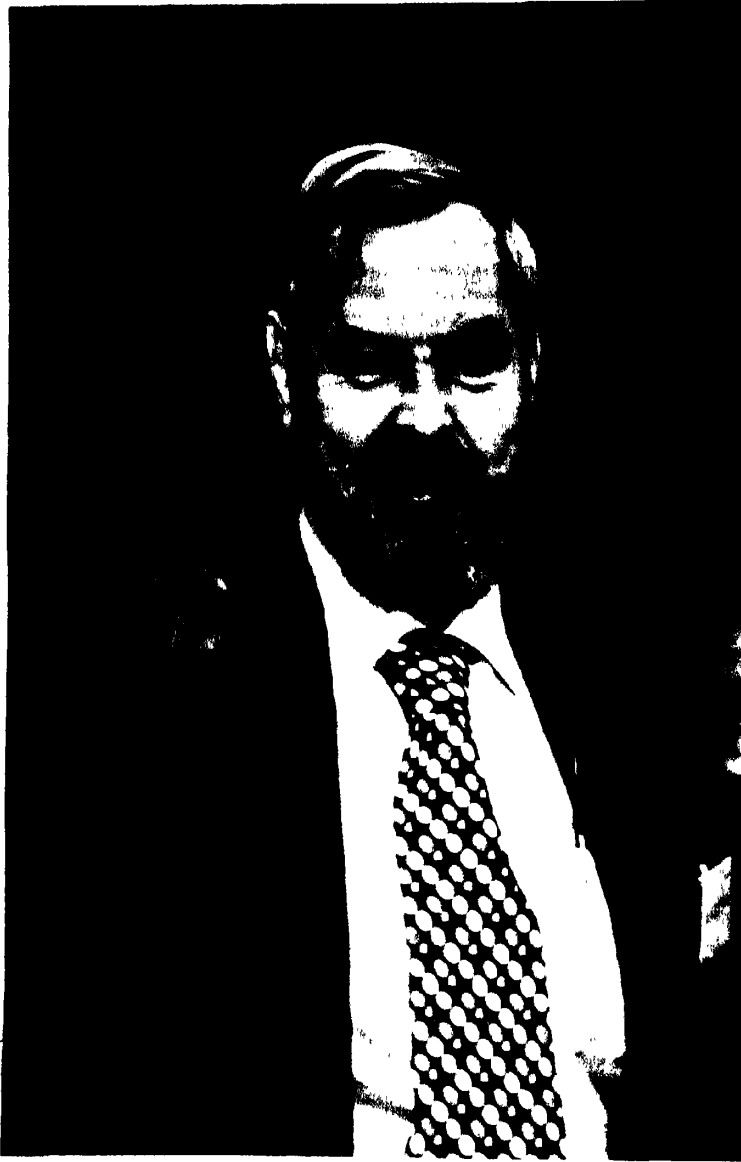
Harry G. Johnson, whose recent untimely death prevented his receiving this award in person, was a consummate professional economist. To him, "the profession" meant a great deal, far more than to most of our colleagues; indeed, his life and work were dedicated to it.

For Harry Johnson, economics was a mature, living social science. It was mature in the sense that it grew out of a grand tradition forged by some of the greatest minds of earlier generations; it was a social science in that it dealt with people and institutions in the real world It was alive in the sense that it had much to say to people and to governments about the current problems and decisions that they faced

Harry Johnson's career reflected his devotion to the profession in all of these senses. Conscious of our professional heritage, he took pains to keep it current, integrating new contributions with it; demonstrating again and again how its lessons had meaning for interpreting events and for guiding policy choices. He nurtured the grand tradition of economics, too, in his celebrated role as critic, dealing harshly with those who treated our heritage lightly, or who mistook novel twists for fundamental challenges to the corpus of economic theory.

Intermingled with Johnson's continuing effort to integrate the new with the old, and to provide the profession with a current version of its distilled wisdom, were many substantive contributions to our knowledge. He was a major contributor to the modern theory of tariffs, to the theory of income distribution, and to the integration of the theories of international trade and economic growth. In an entirely different vein, especially in his classic articles "The General Theory after Twenty-Five Years" and "A Survey of Monetary Theory," he clarified the contributions of Keynesian economics and helped move the profession toward a modern synthesis of the Keynesian and monetary approaches. In yet another direction was his work on the monetary approach to the balance of payments. This, the most important innovation in international economic theory of recent years, was foreshadowed by Johnson during the 1950's and was further developed by him and other members of the International Trade Workshop at Chicago during the last years of his life.

These are necessarily only samples of the many contributions contained in the more than 500 scientific papers and twenty books authored by Johnson. Perhaps the best indices of the number and merit of these contributions are those given by the profession itself· the literally thousands of citations to Johnson's works, and the great frequency with which his papers were anthologized.

For Harry G. Johnson's substantive contributions, and for his valued role as critic of and mentor to the profession, we are pleased to recognize his distinction.

HARRY G. JOHNSON

1923–77

THE AMERICAN ECONOMIC REVIEW

September 1978

VOLUME 68, NUMBER 4

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript, $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Articles

Shorter Papers

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style

A Please use the following categories

1 Deaths
2 Retirements
3 Foreign Scholars (visiting the USA or Canada)
4 Promotions
5 Administrative Appointments
6 New Appointments
7 Leaves for Special Appointments (NOT Sabbaticals)
8 Resignations
9 Miscellaneous

B Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment her new title (if any), and the date at which the change will occur

C Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows *March*, November 1, *June*, February 1, *September*, May 1; *December*, August 1

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912

American Indian Relative Ranching Efficiency

By RONALD L. TROSPER*

Although there have been many hypotheses advanced to explain American Indian poverty on reservations, few have been subjected to statistical tests. Using a 1967 sample of Indian and non-Indian ranchers on the Northern Great Plains, this paper examines several competing explanations. Because the sample contains data of individuals' actions, it cannot be used to test explanations about tribal economic decision making. Using profit function tests developed by Lawrence Lau and Pan Yotopolous, one finds that Indian ranchers on the Northern Cheyenne Reservation in Montana and neighboring whites profit maximize to the same degree. The lower level of output per acre for Indians is due to less capital per acre. Indians may face a capital market constraint and may have higher technical efficiency than whites. Both possibilities are related to the higher proportion of leased land among Indians. Since most policy recommendations are based upon a belief that Indians need management advice or capital, the results of this paper suggest a reinterpretation of Indian ranching. Ranching is important because approximately 60 percent of Indian lands in the United States excluding Alaska and Hawaii are open grazing land (see Henry Hough, p. 76).

Three types of explanations have been combined in various ways in the literature on Indian poverty. The first is that Indians have different constraints on their participation in the market than do non-Indians (see Gary Becker, Peter Dorner, Sol Tax and Sam Stanley, James Fitch, Russel Barsh and James Youngblood Henderson, the author, and the American Indian Policy Review Commission). For example, the Bureau of Indian Affairs (*BIA*) approves and supervises all leases of Indian lands, whether owned individually or tribally. Such activity may raise the costs of using Indian land above the rental fees charged. In the capital market, only federal and tribal courts have jurisdiction over civil matters on Indian lands in some states. The resulting uncertainty to banks accustomed to state law may raise the cost of capital to Indians (see John Mudd).

A second explanation is that Indians lack managerial and technical knowledge. Alan Sorkin (p. 67) and Sar Levitan and William Johnston (p. 21) cite weak evidence given by *BIA* land officers that Indians fail to obtain high enough calf yields or choose to raise cattle when sheep would be more profitable. William Brophy and Sophie Aberle (pp. 80–85) appear to support this view, with little evidence.

A third explanation is that Indians have different goals than non-Indians (Robert Bennett, Robert Bigart). Brophy and Aberle (pp. 84–85) repeat the claim of Harry Getty (1961–62, 1963) that the small scale of ranches on the San Carlos Apache Reservation is caused by Indian values. Examples are Indian folkways, kin group patterns, generosity, and sporadic activity. This hypothesis is similar to that of the "inert peasant" or "satisficing peasant" in the economic development theory literature (see Lloyd Reynolds, p. 5).

These three types of explanation are not mutually exclusive. In his survey in 1959, Dorner attributes some effect to all three in explaining Indian poverty. Fitch also considers a combination to be important. Problems in one category might exacerbate difficulties in another dimension. Some may think that Indian managerial knowledge is low because of cultural history—Indians were hunters, not businessmen—while others may think that Indian managerial

*Assistant professor of economics, Boston College. Summer research support from a University of Washington Faculty Scholarly Development Grant was greatly appreciated. An earlier version of this paper was presented at the Eastern Economic Association meetings on April 15, 1977. Comments by James Murray and other participants were very helpful. Remaining errors are my responsibility

knowledge is low because access to education has been withheld. (Compare Gordon MacGregor to Edgar Cahn, pp. 25–54.) Another common belief is that receipt of government rations has changed Indian behavior patterns, making them less self-reliant (see MacGregor, p. 62; Kenneth Boulding, p. 100). Another interaction could occur if goals such as the preservation of tribal governments require that state power to tax and to regulate be limited. A price for utilizing state civil law in the capital market might be that a tribe submit to state jurisdiction in all matters.

Emphasizing the interrelationships may obscure the point that the three explanations are not necessarily related. It is possible that only one applies in a particular case. Nor does one problem necessarily create another. Take, for instance, diffusion of technical knowledge. Many believe government interference or cultural differences prevents Indians from obtaining such knowledge. This argument is not necessarily true and may not be plausible. Suppose that interference slows the rate of diffusion by some percentage. Indians would not be the first to learn of a new vaccination, but they would eventually do so. Such ideas come in spurts. Government interference might double the time required; a move from five to ten years, say. Eventually, Indians would catch up with neighboring non-Indians after each spurt of ideas. Since ranching on the Northern Great Plains has existed since the late nineteenth century, Indians have had time to catch up with neighboring whites. Checking the data will show if they have or not.

Similarly, there is no necessary conflict between values and the efficient use of resources. The theory of the "optimizing" peasant is now ascendant in economic development theory because more facts are consistent with it (see Reynolds, pp. 5–8). People of diverse cultures make allocation decisions similarly. Indians probably are optimizers also. Lorraine Ruffing showed that a community of Navajos responded in the right direction, in that case transferring resources from sheep to cattle raising. In contrast to the *BIA* view cited above, Ruffing found cattle to be more profitable than sheep. She argues that there is no negative relationship between "communal" values and efficient resource allocation.

Two thorough studies of ranching on the Crow and Blackfeet reservations do not assert that Indians are less efficient than non-Indians. Both Ralph Ward et al. and Sidney Tietema, Ward, and Baker stress the importance of many barriers besides ability and values. In generating recommendations based on the studies, Tietema assumed without caveat that Indians would apply average managerial ability to the alternatives which he costed (p. 11).

In addition to furthering our understanding of the position of Indians in the United States, tests of these competing views are important because they have been used to justify different policy recommendations. An incorrect diagnosis supports policies which are either ineffective or which make the real cause worse. Assume that the federal goal is to raise the income of Indians. If high capital costs are the main obstacle, and Indians are already as knowledgeable as other ranchers, then extension expenditures directed at techniques will have little impact. Yet a belief that Indian management is the problem suggests extension expenditures. Should the real problem be *BIA* interference with Indian management, neither a credit program nor an extension program will have an effect; *BIA* administration of either might strengthen *BIA* power.

Current federal policy is based primarily upon a belief that Indian managerial and technical knowledge is the effective constraint. Barsh and Henderson (pp. 309–13) call this belief "leadership-deficit thinking" when applied to tribal decisions. The Indian Self-Determination and Education Assistance Act became law in January, 1975 (Public Law 93-638). It was praised as watershed legislation in both Congress and the Executive Branch. The self-determination portion of the law seeks to strengthen Indian ability to manage tribal affairs. The education assistance portion supports training individuals. The results of this paper suggest that the Northern Cheyenne are al-

ready as good as or better at ranching as are neighboring non-Indians. After controlling for capital and land holdings, any significant difference between Indians and whites favors this group of Indians. Although this result affects only the education assistance portion of the legislation, it calls into question the assumption underlying the self-determination part as well.

Firm-level data should serve to exclude some of the explanations. In order to do so, we must first derive implications from the theories about the comparative behavior of Indian and non-Indian ranchers. If Indians and whites differ in abilities, constraints, or goals, they will operate ranches at different scales, factor input ratios, levels of technical efficiency, and levels of profit. The task is to disentangle the various effects. Rather than attempt to develop each theory independently and then apply the results to the data, I will proceed directly to discussing each of the effects after describing the data base. The Appendix contains the definitions of the variable notation.

I. The Data

In the process of designing a study of Indian livestock operators on the Northern Cheyenne Reservation, members of a research office in the *BIA*, the Missouri River Basin Investigations Project (*MRBIP*) discovered that the Economic Research Service of the Department of Agriculture was simultaneously undertaking a similar study. The *BIA* therefore collected comparable data for the same year, 1967. Both agencies wished to determine reasonable lease rates for grazing land. The *BIA*, in addition, wished to discover what could be done to aid Indian livestock operators. Upon reading the resulting comparative study (see U.S. Department of the Interior), I requested and obtained access to the original data, with names deleted to protect individual privacy.[1]

The non-Indian ranches are a stratified random sample of ranchers with more than fifty head of cattle located in southeastern Montana, northwestern South Dakota, and northeastern Wyoming. The Northern Cheyenne Reservation is located in southeastern Montana. The Indian sample is all Indian operators with more than ten head who wished to cooperate. There may have been some bias resulting from self-selection of the Indian ranchers; however, the resulting forty-three ranches used here are approximately half the total number of Indian ranchers on the reservation.[2]

The collected information is fairly complete. Table 1 provides descriptive statistics. The value of sales, the type of cattle sold (primarily calves), the value of capital (equipment, buildings and improvements, initial cattle inventory), the quantity of land owned and leased, and the amount and cost of hired labor are available for each ranch. Family labor data are weak; the *BIA* interviewer did not obtain a careful estimate of time actually worked by operators who said they worked "full time", the data on whites are similarly suspect. Those ranches which did not hire labor did not report what they would have had to pay. In order to preserve the size of the sample, wage data for such ranches had to be provided. The wages paid by other ranches in the same county were used for the non-Indian ranches. Indian ranches were assigned the average wage paid by those which hired labor. "Wage paid" includes board and room, if provided. Although these substitutions artificially reduce the variation in the wage data, the difference in hired labor costs is significant in the sample of ranches which hired labor.

The cost of borrowing is unavailable for either group. Education of operator is available only for Indians; the median education achieved by Indian operators is 10.8 years. According to the U.S. Bureau of the

[1]I would like to thank the Planning Support Group (formerly the *MRBIP*) of the Bureau of Indian Affairs and the Economic Research Service of the Department of Agriculture for providing access to raw data.

[2]There were a total of ninety-seven Indians either ranching or farming or both on the Northern Cheyenne Reservation in 1967. The *MRBIP* report does not give the percentage sampled by the Department of Agriculture.

TABLE 1 - DESCRIPTIVE STATISTICS

	Full Sample		Indians		Whites	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Number of Observations	113	-	43		70	-
Capital ($ Value)	83,800	49,600	51,800	27,600	103,000[e]	49,800
Land (Acres)[a]	7,480	4,910	6,390	3,860	8,140[c]	5,350
Land ($ Value)	226,000	140,000	185,000	109,000	252,000[d]	150,000
Rental per Acre ($ Per Year)	.675	468	.721	.173	.646	.577
Owned Land Value/Operated Land Value	.450	376	.0668	.0809	696[e]	.244
Labor (Weeks/Year)	81 7	38 3	60.9	29 1	94.5[e]	37.7
Daily Wage[b]	10.4	1 58	11.4	1 77	9.69[e]	.936
Production ($)	23,900	17,700	16,200	12,500	28,600[e]	18,800
Average Value per Animal Sold	126	37 6	111.	31.7	136.[c]	37 8
Production Less the Cost of Hired Labor	23,100	16,700	15,700	11,900	27,600[e]	17,600

Source: *MRBIP* and Economic Research Service.
[a] Both Indian and white ranches have the same median size, 6780 acres.
[b] The statistical tests describe data for those ranches which hired labor
[c] White and Indian means different at the 5 percent level
[d] White and Indian means different at the 1 percent level
[e] White and Indian means different at the 05 percent level

Census (1973), all Indians twenty-five years old and over on the Northern Cheyenne Reservation had a median of nine years of education in 1970. Rural Montana farm residents in 1970 in the same age class in the counties from which the non-Indian sample was drawn had medians slightly above twelve years of school completed (U S Bureau of the Census, 1971).

The *BIA* study provides average values for land of each type (see U.S. Department of the Interior, p. 60), land value for each ranch was estimated using the following values: $125 per acre for irrigated land, $60 per acre for dry cropland, and $25 per acre for range This procedure creates error by assuming values are constant across groups within each type of land Market data on value would be biased as a measure of productivity, however, because legal restrictions on the sale of Indian land depress its market value.

All but one rancher leased at least some grazing land. Indians leased over 90 percent of their land, measured by value rather than acres, while the non-Indians leased approximately 30 percent of theirs. Reported lease rates in Table 1 refer to the average on leased land.

In summary, the data are sufficiently complete for comparative study of Indians and whites. None of the data gaps are insurmountable. Cross-section data serves well for examining relative efficiency. I now turn to the comparisons

II. Scale and Factor Ratios

Observations about the scale of enterprise cannot distinguish among the explanations because each of them either suggests that Indian ranches will be smaller on average than non-Indian ranches, which is the case, or makes no prediction regarding scale. Examination of factor input ratios is more productive.

Regarding scale, if Indians face higher costs for some inputs, the optimum size of operation will be smaller than for whites, who have access to cheaper inputs. Second, with constant returns to scale, better managers, having more ability (a factor input), will operate at a larger scale. Firm size will be determined by the owner's knowledge. Finally, if Indians are not profit maximizers, they will not have an incentive to operate at the profit-maximizing level. Their firm size could be above or below the profit-

TABLE 2—COMPARISON OF RATIOS

Variable	Means Indians	Means Whites	Ratio of Means
Production /Land Value	.0838	.121[a]	.691
Capital /Land Value	.297	.459[a]	.647
Labor /Land Value	.000401	.000469[b]	.85
Capital/Labor	941	1098[b]	.86

Source: See Table 1

[a] Difference in means significant at .05 percent level

[b] Difference not significant at 5 percent level.

maximizing size; thus there is no prediction from the view that Indians are not profit maximizers.

Examination of factor input ratios provides some differentiation and is consistent with the idea that Indians face higher capital costs. High capital costs and profit-maximizing behavior imply that the capital-labor and capital-land ratios will be lower for Indians than for non-Indians. Neither lack of managerial ability nor nonprofiting-maximizing behavior suggests anything specific about such input ratios. A management error could cause input ratios to be wrong in either direction. If the mean of each group is the correct ratio, then the group with the larger variance will be making more errors than the other group. A test against the null hypothesis that input ratio variances are equal is appropriate; the test shows one cannot reject the hypothesis.[3]

Table 2 gives the capital-land, capital-labor, and labor-land ratios for the ranches by group. For each of the three, the white ranchers have the higher ratios. The difference is significant only for the capital-land ratio, when land is measured by value. If both groups are responding to actual prices, this evidence suggests Indians face higher capital costs because rental rates are not significantly different between the groups. Since Indians pay significantly higher wages, one cannot examine the capital-labor ratio to test for higher capital costs.

[3]The test statistic comes from Alexander Mood and Franklin Graybill, pp. 307–08.

Although the test on capital-land ratios suggests Indian access to capital is inhibited compared to whites, the source of the problem may not be that suggested in the literature: lack of access to credit due to tribal rather than state jurisdiction.[4] Northern Cheyenne ranchers owned an average 6.7 percent of their operated land, while whites owned 69.6 percent of their operated land. Greater cash flow from owned land is a potential source of capital. Insecurity of tenure on leased land may discourage investment. The following equation shows there is a significant relationship between the ratio of capital to land, R, and the proportion of operated land which is owned, S. Land is measured by value rather than acres.

$$(1) \qquad R = \underset{(.021)}{.290} + \underset{(.035)}{.236}\, S$$

$R^2 = .29$; Number of observations = 113

Standard errors are in parentheses. A Chow test shows the relationship is the same for both groups. Similarly, if a dummy variable for whites is added to equation (1), a t-test shows its coefficient is not significantly different from zero.[5]

The difference between the means of the capital-land ratio in Table 2 is .162. The difference between the means of the proportion of owned land is .629; the product of this difference and the coefficient of S in

[4]A capital market constraint has been suggested by many (see Dorner, Mudd, Frank LaFontaine, Barsh and Henderson, and the author). These authors identify two separate legal situations. 1) Federal law often explicitly prohibits mortgages of trust land and assets generated from that land, whether owned by individuals or by a tribe (for examples see LaFontaine, pp. 30–34). Such land or assets are "restricted" and held in "trust." 2) For unrestricted individual or tribal property such as cattle, civil jurisdiction, with some exceptions, is in the tribal court (see LaFontaine, p. 31). The first condition applies to individual land even under tribal law. Unable to mortgage their land, Indians can offer no security for loans. This is a cost to them.

[5]The full equation, with D^W the dummy variable for whites, is as follows:

$$R = \underset{(.021)}{.285} + \underset{(.065)}{.185}S + \underset{(.048)}{.046}D^W$$

$R^2 = .29$, Number of observations = 113

equation (1) is .148. Therefore, nearly all of the difference in the means of the capital-land ratio can be accounted for by differences in land ownership. Capital ownership is significantly related to land tenure and some discussion of possible causes follows.[6]

It is difficult for individual Indians to assemble large tracts of land under one owner. In 1967, about 60 percent of the grazing land on the Northern Cheyenne Reservation was tribal land which individuals cannot purchase. Individual Indians or their heirs owned the remainder in small "allotments" averaging 150 acres each. These allotments were created when the allotment policy was applied to the reservation in 1932 Beginning in 1887, successive Indian reservations were allotted in an attempt by the federal government to break up reservations; assigning individual property to Indians was to lead to eventual assimilation and disappearance of tribes (Harold Fey and D'Arcy McNickle, pp 79-90). The policy ended in 1934, but allotments remain. Purchase of allotments is made difficult by a complicated heirship situation which created multiple ownership for the majority of allotments. In these cases, all owners must agree on a sale; since many heirs are minors, they need court-appointed guardians to approve a sale. Sales are therefore difficult to arrange (see U.S. Congress, p. 348). Since individual ownership of the proper size ranch for a family operation is difficult, land use is instead governed by range leases administered by the *BIA*. Under this situation, it is not clear that an Indian rancher could purchase sufficient reservation land even with credit.

Although most white ranchers own a greater share of their land than do Northern Cheyenne ranchers, the statistical relationship between the owned proportion and the capital-land ratio is the same for both This suggests but does not prove both groups are governed by the same capital access rules even though whites do not face the jurisdictional or land-tenure problems alleged to trouble Indians. One possible explanation is that insecure tenure based on leases, whether Indian or white, discourages capital investment; a credit constraint is not effective. Another possible explanation could lie in the time path of investment on each ranch, although no dynamic model is attempted here. Land use and ownership on the reservation was changed in 1932 and most Indian ranches have been established since then. Many of the white ranchers inherited substantial land acquired by their parents or grandparents in the late nineteenth or early twentieth centuries when the cost of assembling large tracts was low. Cash flow from the owned land could aid investment and Indians started late. Cross-section data from one year is insufficient to settle these issues.

In summary, although Indians have lower capital-land ratios which correlate with land tenure, identifying causation is difficult. The lower average size of Indian ranches may also be related to the age of the ranch as a family enterprise.

III. Output-Factor Input Ratios

Tests on scale and factor proportions are consistent with the view that Indians have higher capital costs, but they cannot rule out the possibilities that Indians have less managerial knowledge or different goals. Even though the white and Indian ranchers act as if they face different factor prices, if one assumes that they operate under the same production function, it should be possible to test whether or not they are equally efficient As emphasized by Lau and Yotopolous, efficiency has two components. "Technical" efficiency refers to a scale parameter in the production function. "Price" efficiency describes maximizing behavior. To illustrate technical efficiency, assume a fixed form for the production function $G(N, T, K)$ with the arguments labor, land, and capital, respectivey. Define two production functions which differ by a scale parameter:

$$(2) \qquad F^I = A^I G; \qquad F^W = A^W G$$

[6] I thank James Murray for suggesting the importance of land tenure

Because of the econometrics involved, it is convenient to test for the equality of A^I and A^W using profit rather than production functions. Before doing so, however, important insights can be gained by examining ratios of output to factor input.

Sorkin has argued that Indians are less efficient than whites because they produce less per farm and per acre (p. 67). This argument is incorrect. Indians use different amounts of capital and labor per acre. Since Indians have less capital per acre, one would expect them to have less output per acre. One must compare the differential: do Indians produce less than what would be expected, given that they use less capital per acre?

In terms of the production functions, we transform them by dividing output and the other two inputs by the amount of land. Assuming constant returns to scale, then

$$(3) \qquad Y^i/T^i = A^i G(N^i/T^i, 1, K^i/T^i)$$

with the index i running over I and W. If the proportions Y/T, K/T, and N/T are approximately the same for the two groups, then we would conclude that $A^I = A^W$. If the three ratios are not the same, then we must assume something more specific about the form of G and take into account factor prices in order to make the test correctly. Examining the ratios has the virtue of simplicity.

Table 2 presents this test. The results appear to confirm the view that Indians and whites have the same technical efficiency. Although Indians produce 69 percent as much per dollar of land value, they have 65 percent as much capital per dollar of land and 85 percent as much labor per dollar of land. The 85 percent figure may be too high because Indian family labor input may be overestimated.

IV. Relative Technical and Price Efficiency

The fact that the input ratios are not exactly the same for Indians and whites suggests that more sophisticated tests be used. Adopting a specific form for the production function allows correction for the fact that factor inputs vary across groups. Price inefficiency creates an identification problem for tests of technical efficiency. As Yotopolous and Lau show, one should simultaneously test for both types of efficiency if price inefficiency is suspected. Their method uses a Cobb-Douglas variable profit function. Since their paper derives the method, I will explain only the procedure. The important tools are a definition of price efficiency and a variable profit function.

Price efficiency is defined as follows. Suppose that the two types of firms respond to factor prices differently. In the labor market, for instance, rather than equating the value of the marginal product to the wage rate, each firm may interpose a proportionality term k^i, as follows:

$$(4) \qquad \partial F^i(N,T,K)/\partial N = k^i w'^i/P^i$$

where w'^i = wage rate for firm i
P^i = output price for firm i
$i \in (I, W)$

If $k^I = k^W = 1$, both firms have absolute price efficiency. If $k^I = k^W \neq 1$, both firms are not price efficient but have the same relative price efficiency. If $k^I \neq k^W$, the firms have different price efficiency.

Although the two firms may differ in technical constants, A^I and A^W, to introduce the profit function, let us write a common Cobb-Douglas production function as follows:

$$(5) \qquad G(N,T,K) = A\, N^{\alpha} T^{\beta_1} K^{\beta_2}$$

where A = technical constant
N = labor
T = land
K = capital
α = elasticity of output with respect to labor
β_1 = elasticity of output with respect to land
β_2 = elasticity of output with respect to capital

The data describe activities in one season. In this period, capital and land are fixed and labor is variable. Define normalized

variable profit π as current revenue less current variable costs, in this case the wage bill, divided by the price of output. Duality theory shows that there is a variable profit function corresponding to the production function defined by (4) as follows:

$$(6) \quad \pi = H(w,T,K) = A^{(1-\alpha)^{-1}}(1-\alpha)^{-a}w^{a}T^{b_1}K^{b_2}$$

where π = normalized variable profit
$w = w'/P$ = normalized wage rate
$a = -\alpha(1-\alpha)^{-1}$
$b_1 = \beta_1(1-\alpha)^{-1}$
$b_2 = \beta_2(1-\alpha)^{-1}$

Variable profit is a decreasing function of the normalized wage rate and an increasing function of land and capital. Variable profit depends upon the wage rate and fixed factors; these are exogenous variables. A simultaneous equation model is not required to estimate parameters, those in the profit function are functions of those in the production function. Further, because Northern Cheyenne and neighboring whites differ in levels of capital and land, this technique allows us to correct for these differences directly.

If the firms vary in technical efficiency, one can substitute either A^I or A^W for A in (5). If the two firms vary in their price efficiency, however, this straightforward substitution is not possible. To incorporate k^I and k^W as defined in (4) above, Lau and Yotopoulos show that the observed variable profit function is

$$(7) \quad \pi^i = A^i_* w^a T^{b_1} K^{b_2}$$

where, for $i \in (I, W)$,

$$(8) \quad A^i_* = A^{i(1-\alpha)^{-1}}(1-\alpha/k^i)(k^i)^{-a}\alpha^{-a}$$

The goal is to test whether or not $A^I = A^W$ by using a dummy variable to estimate A^W_*/A^I_*. The test will not work if $k^I \neq k^W$, because the ratio A^W_*/A^I_* involves the constants k^I and k^W, as follows:

$$(9) \quad \frac{A^W_*}{A^I_*} = \left(\frac{A^W}{A^I}\right)^{(1-\alpha)^{-1}}\left(\frac{1-\alpha/k^W}{1-\alpha/k^I}\right)\left(\frac{k^W}{k^I}\right)^a$$

If $k^I = k^W$, however, the ratio simplifies to the following:

$$(10) \quad \frac{A^W_*}{A^I_*} = [A^W/A^I]^{(1-\alpha)^{-1}}$$

First we must test $k^I = k^W$. To do so, Yotopoulos and Lau exploit a convenient property of the Cobb-Douglas function. When factor prices vary, quantities adjust to keep factor shares constant. The same is true for the ratio of variable factor costs to variable profit. Suppose a firm responds efficiently to the wage rate, i.e., $k^i = 1$. Then

$$(11) \quad -wN/\pi = -\alpha(1-\alpha)^{-1} = a$$

But if a firm is acting as if the wage rate is k^iw, then the ratio k^iwN/π will be constant as w varies, not wN/π. We do not observe k^iw. Yotopoulos and Lau show that when $k^i \neq 1$,

$$(12) \quad -wN/\pi = a(1-\alpha)[k^i(1-\alpha/k^i)]^{-1} = a^i$$

This equation defines a^i. We can therefore test the hypothesis $k^I = k^W$ by testing whether or not the ratio of the observed wage bill to variable profit is significantly different for Indians and whites The appropriate equation is

$$(13) \quad -wL/\pi = a^I D^I + a^W D^W$$

where D^I = 1 for Indians, 0 for whites
D^W = 1 for whites, 0 for Indians
L = hired labor

We substitute hired labor L, for total labor N, for this sample because family labor data are weak. Family labor is omitted and treated as fixed.[7]

If the test shows equal relative price efficiency across groups, we can proceed to test absolute price efficiency and relative technical efficiency with the Cobb-Douglas profit function. The estimating equation uses a dummy variable for whites. The two profit functions are

$$(14) \quad \pi^I = A^I_* w^a T^{b_1} K^{b_2}$$

$$(15) \quad \pi^W = A^I_*(A^W_*/A^I_*)w^a T^{b_1} K^{b_2}$$

[7]There is no significant difference between the two groups in the proportion of total labor which is hired.

TABLE 3—ESTIMATION OF COBB-DOUGLAS PROFIT AND LABOR DEMAND FUNCTIONS, LABOR VARIABLE

	Model 1			Model 2	
	Ordinary Least Squares	Zellner's Method			
Parameters		Unrestricted	2 Restrictions[a]	2 Restrictions[a]	3 Restrictions[b]
Variable Profit Function					
$ln\, A^I_*$	−6.98	−7.59	−7.62		-
	(.75)	(.71)	(.65)	-	-
d	−.185	−.229	−.229	-	-
	(.11)	(.10)	(.091)	-	-
$ln\, B^I_*$	-	-	-	−7.51	−7 42
	-	-	-	(.45)	(.44)
c	-	-	-	−.273	−.340
	-	-	-	(.12)	(.081)
f	-	-	-	− .0703	-
	-	-	-	(.095)	-
a	0779	−.00393	− .0301	− .0301	− .0301
	(.28)	(.26)	(0044)	(0031)	(.0031)
b_1	.178	.197	.198	160	.166
	(.11)	(11)	(10)	(074)	(074)
b_2	1 02	1 05	1 05	1 08	1.06
	(.12)	(11)	(.11)	(080)	(077)
Labor Demand Function					
a	− 0291	− 0291	− .0301	− .0301	− 301
	(0072)	(0072)	(.0044)	(0031)	(0031)
a^W	− 0307	− 0307	− 3001	− 0301	− 0301
	(0056)	(.0056)	(.0044)	(0031)	(0031)

Source See Table 1

Notes Numbers in parentheses are standard errors Notation is defined in the Appendix Estimating equations are as follows

Model 1

$$ln\, \pi = ln\, A^I_* + dD^W + a\, ln\, w + b_1\, ln\, T + b_2\, ln\, K$$

$$-wL/\pi = a^I D^I + a^W D^W$$

Model 2

$$ln\, \pi = ln\, B^I_* + cS + f\, D^W + a\, ln\, w + b_1\, ln\, T + b_2\, ln\, K$$

$$-wL/\pi = a$$

[a]The restrictions are $a^I = a$ and $a^W = a$

[b]The restrictions are $a^I = a, a^W = a$, and $f = 0$

Taking logarithms, the following equation tests relative technical efficiency:

$$(16) \quad ln\, \pi = ln\, A^I_* + d\, D^W + a\, ln\, w + b_1\, ln\, T + b_2\, ln\, K$$

$$\text{where} \quad d = ln\,(A^W_*/A^I_*) = (1/1 - \alpha)\, ln\,(A^W/A^I)$$

If d is significantly different from zero, then A^W and A^I are not equal. We can also test absolute price efficiency because if $k^I = k^W = 1$, then $a^I = a^W = a$, as can be seen by inspecting the equation defining a^I in (12).

The estimating equations are (13) and (16). The first three columns in Table 3 give the coefficients of these equations. This is Model 1 in that table; a second model is presented shortly. Before examining the results, a few notes on the specification are in order. The average selling price for calves was 30¢ per pound for whites and 29.1¢ per pound for Indians (see U.S. Department of the Interior, p. 51). These prices were used to normalize variable profits and factor

TABLE 4 - TESTS OF HYPOTHESES

Maintained Hypothesis	Hypothesis	Chi-Square Tests: Computed Value	Critical Values: 5 Percent	Critical Values: 1 Percent
		Model 1		
	$a^I = a^W = a$	.04	5.99	9.21
$a^I = a^W = a$	$d = 0$	6 18	3.84	6 64
		Model 2		
$a^I = a^W = a$	$c = 0$	8 41	3.84	6.64
$a^I = a^W = a$	$f = 0$	.27	3.84	6.64

Source: See Table 1. Computed values are $-2\ ln\ \lambda$, where λ is the ratio of the likelihood of the restricted system to the likelihood of the unrestricted system.

prices.[8] Variable profit is the value of production less hired labor costs.

Two estimation techniques are used, ordinary least squares in the first column of Table 3 and Zellner's minimum distance estimator in the remaining columns. Zellner's method is appropriate because there is a restriction across equations and it is reasonable to assume errors in each equation are independent across ranches. The statistical test for the significance of a restriction on the system is $-2(ln\ \lambda)$, where λ is the ratio of the likelihood of the restricted system to the likelihood of the unrestricted system. The test statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of restrictions. Table 4 reports these tests.

The fit of the equations is decent. The $R^2 = .80$ for the profit equation. The standard error on the capital coefficient is very small; the fit is less good for the value of land. The wage variable is not significant, although the sign of the coefficient is negative, as it should be, when estimated with Zellner's method. The poor behavior of the wage variable may be due to its low variation.

We test first the combined hypothesis of equal relative and absolute price efficiency. When the system is restricted to have $a^I = a^W = a$, the value of the likelihood function hardly falls. We cannot reject equal relative and absolute price efficiency. One can see by examining the labor share equation in the unrestricted case that there is little difference between whites and Indians. A t-test for such difference in the equation estimated by ordinary least squares also shows that the hypothesis $a^I = a^W$ cannot be rejected. The test of absolute price efficiency is not very convincing because the standard error of the labor coefficient is so large in the profit equation.

Assuming $k^I = k^W$, a test on the coefficient of the dummy variable for whites in the profit function is a test of relative technical efficiency alone. Using the likelihood ratio test, the hypothesis that $d = 0$ is rejected at the 5 percent level but not at the 1 percent level. A t-test on the coefficient of D^W in the profit function estimated with ordinary least squares gives the same result. In both cases, the coefficient is negative. Whites appear to have lower technical efficiency.[9]

The result surprises those who believe Indians to be equally or less knowledgeable than whites. It is particularly surprising because we have not controlled for education. In this test, land and capital have been

[8]Selling price was not available for each ranch, the figures in the text are averages for those ranches which supplied a price. The tests in Table 4 are the same if one assumes all ranches had the same selling price.

[9]The *MRBIP* study (p. 49) concluded that Indians had lower operating costs than whites, based on a comparison of average costs in each group. No statistical tests were made, however, to see if the cost difference was significant.

controlled across ranches. Because the average size of ranches and the proportion of owned land varies across groups, we must test to see if these variables matter. A test on size suggests that size is not related to efficiency, although economies of scale are present.[10] The proportion of owned land is significant, however A test is constructed as follows. Suppose the technical constants A_*^I are related to the proportion of leased land S, according to this equation.

$$A_*^I = B_*^I e^{cS} \tag{17}$$

where B_*^I is the technical constant in the profit function for each group after correction for land tenure.

Assuming $k^I = k^W = 1$, the estimating equations then become

$$\begin{aligned} ln\,\pi &= ln\,B_*^I + cS + fD^W \\ &\quad + a\,ln\,w + b_1\,ln\,T + b_2\,ln\,K \\ -wL/\pi &= a \end{aligned} \tag{18}$$

where S = value of owned land/value of operated land

$f = ln(B_*^W/B_*^I)$

The estimated coefficients using Zellner's method are shown in the fourth and fifth columns of Table 3 as Model 2 Test statistics are in Table 4[11] The null hypothesis that $c = 0$ can be rejected at the 1 percent level. Efficiency is negatively related to the proportion of owned land. The hypothesis that $f = 0$ cannot be rejected. Corrected for land tenure, $B_*^I = B_*^W$; Indians and whites have equal relative technical efficiency.

Something connected with tenure affects efficiency. Perhaps leased land is treated more harshly—through overgrazing for instance—than owned land by both Indians and whites. Many will agree that a rancher has less interest in the future productivity of leased land than owned land. It is possible that collinearity between race and land tenure, however, causes a spurious conclusion. Good comparative data on the condition of the land by group would provide a definitive answer; the data are unavailable.

Controlling for land tenure, capital, and land, Indians and whites appear equal in price and technical efficiency. This result is still surprising to those who believe Indians are less knowledgable or have different goals than whites.

Assuming for discussion that other Indian tribes with histories of ranching are approximately equal to their neighbors, where might other studies have gone wrong? Many of them assume small scale means lower efficiency. Others cite particular practices which appear inefficient. But "low calf crops"—a popular example—may well be the efficient technical choice in the situations the Indians face. Getty's 1963 study of the San Carlos Apache has no data on inputs or outputs with which to compare the Apaches with other ranchers. His conclusions about causality, therefore, are unsupported. Rolf Bauer's study of the Papago has the same problem.

V. Policy Implications

Four readily available and perhaps widely read books all recommend provision of additional extension services (see Brophy and Aberle, pp. 92 95, 113-14; Sorkin, pp. 67, 182; Levitan and Hetrick, pp. 132, 135, 206; Levitan and Johnston, pp. 21, 78). The books all stress management deficiencies, rarely distinguishing among Indian tribes, industries, or types of jobs. That the

[10]The test for the effect of size is exactly that of Lau and Yotopoulos, who concluded that small farms were technically more efficient than large farms in India This sample was divided into large and small farms at the median, 6780 acres This median divided the full sample and both subsamples in half Dummy variables were redefined as D^L for large ranches in place of D^W and D^S for small ranches in place of D^I Neither the hypothesis $a^S = a^L = a$ nor $d = 0$ could be rejected at the 5 percent level Computed test values were 5 23 and 2 23, respectively. The test for economies of scale is that $b_1 + b_2 = 1$ This hypothesis is rejected with a computed chi-square of 16.42 The 1 percent critical value is 6 64. Using the last column of Table 3, $b_1 + b_2 = 1.23$, with a standard error of .038 and a t-ratio of 6.12. The standard error for the sum is low because b_1 and b_2 have a negative estimated covariance.

[11]Equations (18) assume that $a^I = a^W = a$. A test of this hypothesis, using the second specification for the technical parameters, gives $-2\ ln\ \lambda$ = 05, the hypothesis cannot be rejected

Northern Cheyenne are equal to neighboring whites in ranching technical efficiency is inconsistent with this prevailing view and challenges a fundamental tenet of paternalistic policies. That the two groups have the same relative price efficiency, with respect to hired labor, is inconsistent with the view that all Indians are not profit maximizers. Exhortation that Northern Cheyenne learn to respond to relative prices is not necessary. If other Indians are similar to the Northern Cheyenne, neither policy deserves the emphasis given it to date.

Other policy implications are more difficult to specify. That a credit problem has held back Northern Cheyenne ranchers receives only ambiguous support Rather, lower capital-land ratios and higher technical efficiency on Northern Cheyenne ranches appear related to the high proportion of leased land These observations may result from insecurity of tenure· unsure that the tribal council or owners of allotments will allow the *BIA* to renew his leases, an Indian rancher is less willing to invest and more willing to overgraze, compared to what he would do if he owned the land

Even if one accepts these inferences, caution is advised regarding policy implications, either for the Northern Cheyenne or for Indians generally. Other variables may intervene. For instance, the Northern Cheyenne Reservation lies on large coal reserves; white ranches in Montana and Wyoming are much less likely to be located over coal. Even in 1967 it may have been proper for Northern Cheyenne to underinvest and overgraze, knowing coal was present and likely to be strip mined eventually

It would be irresponsible to recommend private ownership of Indian ranches based on this limited data because the federal government has created a connection between individual land ownership and eradication of tribal existence. Allotments were created with this in mind; in order to obtain full control of his land, an Indian had to transfer it from tribal and federal jurisdiction to state jurisdiction. He became able to mortgage it, it became subject to state taxes, and the tribe lost jurisdiction. A goal of the allotment policy was to eliminate tribes and tribal government (see Edward H. Spicer, pp. 111–13). But Indians have group goals which their governments serve and have treaty rights which individual Indians hold because they are members of a tribe. Indians have found it difficult to obtain equal treatment under state law when local citizens wish to obtain Indian land. Angie Debo documented the experience of Oklahoma Indians in the early twentieth century. For all these reasons, Indians resist attacks on tribal government and reasonably perceive a recommendation that tribal land be divided up as just such an attack.

Other tribal members beside ranchers have an interest in tribal lands which has to be recognized. If the land is being abused and if ranchers could be aided with additional capital, perhaps a shift in land-use regulations and an expanded credit program would accomplish both ends. A program to aid the sale of heirship allotments to tribal members as well as the tribe would serve a similar purpose. It seems clear the capital problem and leasing are related; therefore more detailed recommendations require more knowledge of specific causal mechanisms.

VI. Summary

The results of this paper suggest land tenure or other institutional problems underlie Indian difficulties attaining the operating scale of whites in ranching. No evidence of management deficiences among Northern Cheyenne were observed. A direct application of the Lau-Yotopoulos tests reveals greater Indian technical efficiency This difference, however, is significantly related to the proportion of owned land. An apparent capital access problem is also related to the proportion of owned land. Further research should investigate the nature of these relationships rather than focus upon Indian abilities and goals.

APPENDIX

T = Value of land operated
K = Value of equipment, buildings, improvements, horses, and cattle at first of the year
N = Total (family and hired) labor in weeks per year
L = Hired labor in days per year
Y = Production, equal to sales plus change in cattle inventory plus family consumption
R = K/T
S = Value of land owned/Value of land operated
I = Superscript indicating Indian
W = Superscript indicating white
i = Superscript index over I and W
A = Technical constant in Cobb-Douglas production function
A_* = Technical constant in Cobb-Douglas normalized variable profit function
α = Elasticity of output with respect to labor
β_1 = Elasticity of output with respect to land
β_2 = Elasticity of output with respect to capital
P = The price of output, the fall calf price in cents per pound
π' = Variable profit, the value of production less the cost of hired labor
π = π'/P, normalized variable profit
w' = Cost of hired labor per day
w = w'/P, normalized wage rate
a = Elasticity of variable profit with respect to the wage rate
b_1 = Elasticity of variable profit with respect to land
b_2 = Elasticity of variable profit with respect to capital
k^i = Proportionality factor to test price efficiency with respect to hired labor
D^i = Dummy variable taking 1 for the appropriate superscript value
d = Coefficient of D^W to estimate $ln(A_*^W / A_*^I)$
B_* = Technical constant in variable profit function corrected for land tenure
c = Coefficient of S
f = Coefficient of D^W to estimate $ln(B_*^W / B_*^I)$

REFERENCES

R. L. Barsh and J. Y. Henderson, "Tribal Administration of Natural Resource Development," *N. Dak. Law Rev.*, Winter 1975, *52*, 307–47.

R. W. Bauer, "The Papago Cattle Economy: Implications for Economic and Community Development in Arid Lands," in William G. McGinnies et al., eds., *Food, Fiber, and the Arid Lands*, Tuscon 1971, 79–102.

Gary Becker, *The Economics of Discrimination*, 2d ed., Chicago 1971.

R. L. Bennett, "Economic Development as a Means of Overcoming Poverty," in U.S. Congress, Joint Economic Committee, *Toward Economic Development for Native American Communities*, 91st Cong., 1st sess., Washington 1969, 102–06.

R. Bigart, "Indian Culture and Industrialization," *Amer. Anthro.*, Oct. 1972, *74*, 1180–88.

Kenneth Boulding, *The Economy of Love and Fear: A Preface to Grants Economics*, Belmont 1973.

William A. Brophy and Sophie D. Aberle, *The Indian: America's Unfinished Business*, Norman 1966.

Edgar S. Cahn, *Our Brother's Keeper: The Indian in White America*, Cleveland 1969.

Angie Debo, *And Still the Waters Run*, Princeton 1940.

P. P. Dorner, "The Economic Position of the American Indians: Their Resources and Potential for Development," unpublished doctoral dissertation, Harvard Univ. 1959.

Harold E. Fey and D'Arcy McNickle, *Indians and Other Americans*, New York 1970.

J. B. Fitch, "Economic Development in a Minority Enclave: The Case of the Yakima Indian Nation, Washington," unpublished doctoral dissertation, Stanford Univ. 1974.

Harry T. Getty, *The San Carlos Apache Cattle Industry*, Tucson 1963.

———, "San Carlos Apache Cattle Industry," *Hum. Org.*, No. 4, 1961-62, *20*, 181-86.

Henry W. Hough, *Development of Indian Resources*, Denver 1967.

F. LaFontaine, "The Native American Credit Problem," *Amer. Indian Law Rev.*, Summer 1974, *2*, 29-40.

L. J. Lau and P. Y. Yotopoulos, "A Test for Relative Efficiency and Application to Indian Agriculture," *Amer. Econ. Rev.*, Mar. 1971, *61*, 94-109.

Sar A. Levitan and Barbara Hetrick, *Big Brother's Indian Programs—With Reservations*, New York 1971.

——— **and William B. Johnston**, *Indian Giving: Federal Programs for Native Americans*, Baltimore; London 1975

G. MacGregor, "Barriers to Economic Development," in U.S. Congress, Joint Economic Committee, *Toward Economic Development for Native American Communities*, 91st Cong., 1st sess., Washington 1969, 61 65.

Alexander M. Mood and Franklin A. Graybill, *Introduction to the Theory of Statistics*, 2d ed., New York 1963.

J. O. Mudd, "Jurisdiction and the Indian Credit Problem," *Mont. Law Rev.*, Summer 1972, *33*, 307 20.

Lloyd G. Reynolds, "Agriculture in Development Theory: An Overview," in his *Agriculture in Development Theory*, New Haven 1975, 1-24.

L. T. Ruffing, "Navajo Economic Development Subject to Cultural Constraints," *Econ. Develop. Cult. Change*, Apr. 1976, *24*, 611-23.

Alan Sorkin, *American Indians and Federal Aid*, Washington 1971.

Edward H. Spicer, *A Short History of the Indians of the United States*, New York 1969.

S. Tax and S. Stanley, "Indian Identity and Economic Development," in U.S. Congress, Joint Economic Committee, *Toward Economic Development for Native American Communities*, 91st Cong., 1st sess., Washington 1969, 75-96.

S. J. Tietema, "Indians in Agriculture: III. Alternatives in Irrigation Farming for the Blackfeet and Crow Indian Reservations," *Mont. Agr. Experim. Sta. Bull.*, No. 542, Bozeman 1958.

———, **R. E. Ward, and C. B. Baker**, "Indians in Agriculture: II. Cattle Ranching on the Blackfeet Reservation," *Mont. Agr. Experm. Stat. Bull.*, No. 532, Bozeman 1957.

R. L. Trosper, "The Economics of Resource Development," in *Handbook of North American Indians*, Vol. 2: *Indians in Contemporary Society*, Washington forthcoming.

R. E. Ward et al., "Indians in Agriculture: I. Cattle Ranching on the Crow Reservation," *Mont. Agr. Experim. Sta. Bull.*, No. 522, Bozeman 1956.

P. A. Yotopoulos and L. J. Lau, "A Test for Relative Economic Efficiency: Some Further Results," *Amer. Econ. Rev.*, Mar. 1973, *63*, 214-23.

A. Zellner, "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *J. Amer. Statist. Assn.*, June 1962, *57*, 348-68.

American Indian Policy Review Commission, Task Force 7, "Reservation and Resource Protection and Development," Final Report, Washington 1976.

U.S. Bureau of the Census, *Census of Population: 1970, General Social and Economic Characteristics, Montana*, Final Report PC(1)-C28, Washington 1971.

———, *Census of Population: 1970*, Subject Reports, *American Indians*, Final Report PC(2)-1F, Washington 1973.

U.S. Congress, Committee on Interior and Insular Affairs, *Indian Heirship Land Survey of the Eighty-Sixth Congress First Session*, Washington 1960.

U.S. Department of Commerce, *Federal and State Indian Reservations and Indian Trust Areas*, Washington 1974.

U.S. Department of the Interior, Bureau of Indian Affairs, Missouri River Basin Investigations Project (*MRBIP*), *Ranch Management: Northern Cheyenne Indian Reservation, Montana*, Report 194, Billings 1969.

Optimal Pricing of Local Telephone Service

By BRIDGER M. MITCHELL*

Although payment for nearly all other goods and services, including toll (long distance) telephone calls, increases with greater consumption, nearly 90 percent of the residential telephone subscribers and more than half the business subscribers in the United States now pay a flat monthly rate for local calls (see Larry Garfinkel).

Recently, however, the telephone companies and regulatory commissions have been moving cautiously toward imposing usage charges for local telephone calls. There is renewed interest in what is currently termed "usage-sensitive pricing" (*USP*). It is due to the combined forces of inflation, increased local usage, and competition from independent firms that sell telephone terminal equipment and supply private toll lines to business customers. Under *USP*, the pricing of such services as telephone installation, directory assistance, and minutes of calling is based on incremental rather than average costs.

Since World War II, technological advances have benefited long distance far more than local telephone calling. Development in microwave communications, coaxial cable, satellites, and waveguides have dramatically lowered the costs of long distance transmission. In contrast, the costs of local service have moved upward since the late 1960's at a rate not far below the general price index (see AT&T, 1975). Faced with a continuing stream of requests for local telephone rate increases, state regulatory commissions are finding the concept of tying prices to usage increasingly attractive.

AT&T and some of the independent telephone carriers are beginning to test *USP* plans in several cities. With flat rate tariffs, increases in local calling add to carrier costs but not to their revenues. The local calling rate per subscriber has increased by an average of 2 percent for the past seven years. AT&T's chairman has stated, "We are moving more and more in the direction of usage-sensitive pricing" (see *Washington Star News*, p. A-12), and according to newspaper accounts of AT&T management documents, the Bell System plans to phase out flat rate telephone service and begin charging for each local call in major metropolitan areas in 1978–80 (see *Seattle Post-Intelligencer*, pp. 1, 10).

Still, regulatory commissions, consumer groups, and the carriers themselves remain cautious about requiring usage-sensitive tariffs for all subscribers. The prospective gains from prices more closely related to costs are at least partially offset by the added costs of metering equipment and billing. Most telephone subscribers prefer flat rates, according to Bell System marketing surveys (see Garfinkel, p. 28). And it is by no means clear which groups of subscribers will be helped and which harmed by such revisions in local tariff structures. Will the poor end up paying more because they use their phones more? Should different prices be charged for calls at different times of day? On what bases should the monthly and per call rates be determined?

The purpose of this paper is to sort out some of the questions of economic efficiency and equity that arise when changes are considered in the methods of pricing local telephone services. In Section I, I construct a

*Department of economics, The Rand Corporation, and the International Institute of Management, Berlin This study was supported under a grant from the John and Mary R. Markle Foundation. I am indebted to Walter S. Baer for numerous contributions to this paper, to S. C. Littlechild and Patricia Munch, and to the managing editor and a referee of this *Review* for critical and constructive review of a draft. I have benefited as well from the comments and suggestions of James H. Alleman, Stanley M. Besen, William S. Comanor, John M. Drew, Leland L. Johnson, John A. Kay, Edward D. Lowry, Carl Pavarini, John Rolph, Ralph Turvey, and Chris Witze. Bryant Mori assisted with the computer programming. See my 1976 paper for a more extensive version of this paper, which also discusses optimal flat rate and optimal peak load tariffs.

model of the demand for local telephone service, distinguishing between the demand for telephone *connections* and the demand for telephone *calls* and incorporating the principal economic characteristics of the market for telephone service. Section II provides a quantitative assessment of the benefits and costs of adopting *USP* for local exchange calling. Data are reviewed on the demand for and cost of local service, and that information is used to calibrate the model of telephone demand to *U.S.* conditions. I then compare flat rate and measured service tariffs in terms of the number of subscribers and the volume of local calls at different levels of consumer income. Finally, the optimal prices are computed for measured service tariffs, and the economic gains such methods of pricing are likely to achieve are assessed. The Appendix summarizes technical details of the calibration of the model and the calculation of optimal tariffs.

I. The Demand for Telephone Service

The market for local telephone service in the United States is characterized by incomplete penetration at existing levels of flat rates and wide variation in the number of calls per month made by subscribers. A welfare analysis of two-part tariffs requires knowledge of demand at different prices, but no data are available as yet on calling rates under usage-sensitive pricing.

To develop a model of the demand for residential telephone service that incorporates these central empirical facts, let us first consider the behavior of a single customer and then the aggregate demand of a population of customers who have similar incomes but differing preferences for telephone service. We subsequently examine the influence of income on demand. A central feature of this model is explication of the link between the demand for telephone calls and for telephone connections. For convenience, "calls" are used as a measure throughout the report, although one could quantify telephone use equally well by using minutes of calling or the "erlang" unit from telephone traffic engineering. Also used interchangeably are the terms "metered" and "measured" service. In practice, metered service applies to the number of calls, measured service to the number of call minutes.

To highlight the principal effects of alternative telephone tariffs, I make a number of simplifying assumptions. My model neglects the influence of toll calls on the demand for service.[1] Also ignored are the dynamic effects of the number of subscribers in the telephone network on the value of service to any one subscriber.[2] Although such externalities may be important in other countries, they should have only a limited marginal effect on demand in a system with a high saturation of subscribers.[3]

A. A Single Consumer

I initially assume that a single consumer has constant marginal utility of income and a separable utility function for telephone calls q and other goods x:

$$U(x,q) = x + V(q) \tag{1}$$

These conditions imply that there is no income effect on the number of calls demanded, an assumption frequently made in applied welfare economics when the effects of price changes are analyzed.[4] Since the magnitude of the price changes that are of policy interest will be small relative to a consumer's income, the quantities derived from this utility function will closely approximate those that will be obtained from a more general formulation. However, the number of *subscribers* cannot be assumed

[1] Toll calls account for about 48 percent of telephone revenue, according to AT&T (1975).

[2] See Lyn Squire or Littlechild (1975). The dynamic role of such consumption externalities in the development of telephone service is explored in Roland Artle and Christian Averous and in Burckhard Von Rabenau and Konrad Stahl.

[3] Present flat rate telephone tariffs are in fact partly based on the number of telephones within a subscriber's local service area. For *U.S.* data relating monthly flat rate and number of telephones, see Paul Behrman and Anthony Oettinger, p. 30a.

[4] This "income effect" is negligible for goods such as telephone service, which are a small fraction of a consumer's budget and do not have unusually large income elasticities of demand.

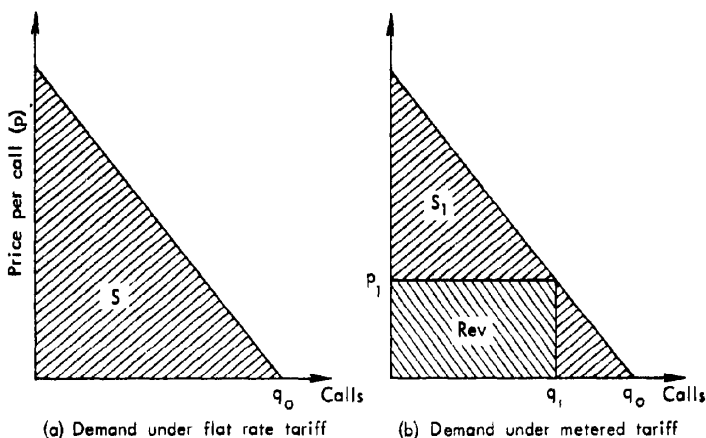



FIGURE 1. A SINGLE CONSUMER'S DEMAND CURVE FOR CALLS

to be independent of the level of income. I will explicitly introduce such a relationship later.

The consumer maximizes utility in equation (1) subject to the budget constraint

$$y = x + \delta(L + pq) \tag{2}$$

where y is income measured in units of x, L is the fixed monthly charge for telephone service, and p is the per call rate (the variable charge).[5] The parameter $\delta = 1$ if the consumer subscribes to telephone service; if he does not, $\delta = 0$.

We can recast the optimization problem as one of maximizing the difference between the utility of telephone service and its cost, that difference being the net economic gain from having a telephone:

$$\max_{\delta, q} \left[V(q) - \delta(L + pq)\right] \tag{3}$$

The solution may be stated as:

Demand function for calls:

$$q = \begin{cases} q(p) & \text{if } \delta = 1 \\ 0 & \text{if } \delta = 0 \end{cases} \tag{4}$$

[5] For my purposes, the one-time "hookup" charge for telephone installation may be annualized and included in the monthly charge L. However, it may be empirically important to distinguish the hookup charge as a separate parameter of the tariff if there are significant differences across consumers in subjective discount rates and in tenure in one residence. See Sharon Black and Peter Tryon.

Demand function for connections:

$$\begin{aligned} \delta &= 1 \text{ if } R(p) \geq L \\ &= 0 \text{ if } R(p) < L \end{aligned} \tag{5}$$

where the consumer's reservation price, or surplus, is given by

$$R(p) = \max_{q} \left[V(q) - pq\right] \tag{6}$$

A single consumer's demand for calls may be approximated by the linear demand curve in Figure 1a. At a zero price per call, the consumer makes q_0 outgoing calls and enjoys a consumer's surplus equal to the shaded triangular area S under the demand curve. The surplus S measures his willingness to pay to have outgoing telephone service. (The added value of being able to receive calls will be considered shortly.)

Under a *flat rate tariff*, the customer must pay L per month for a telephone; he pays no further charges for calls made. The customer will rent a telephone for outgoing calls provided that $L \leq S$. Once he does so, the level of the flat rate charge has no effect on the number of calls he makes.

Under a *metered tariff*, the consumer must pay p_1 per call. This charge reduces his calling to q_1 (Figure 1b) and raises revenue $Rev = p_1 q_1$. The per call charge also reduces his surplus, or willingness to pay to have a telephone, to the area S_1.

Under a *two-part tariff* $[L_1, p_1]$, the con-

sumer must make a monthly payment L_1 and pay p_1 per call. He will therefore subscribe only if $L_1 \leq S_1$. In general, when instituting a per call charge, the telephone company will have to reduce its monthly rental charge if it is to retain all of its original customers.

Of course, consumers also want to receive calls. Even if they made no outgoing calls at all ($q_o = 0$), they would generally be willing to pay at least a small monthly rental for a telephone. We must therefore add this fixed amount to the surplus S or S_1 to determine whether a given consumer will be a subscriber. The amount of surplus also depends on the consumer's income and on the availability and cost of substitutes for a residential telephone. In a later section, I discuss the role of income, but throughout the analysis, it is assumed that the cost and opportunities for using public (coin) and neighbors' telephones are not affected by changes in tariffs.

Assume that individual consumers face the same tariff $[L,p]$ but differ in their preferences, the ith consumer's utility from telephone calls being $V_i(q)$. For a given population of consumers, there will be a statistical distribution of such preferences. Consequently, the market demand curves can be determined by using that distribution to construct a distribution of reservation prices $R_i(p)$. From that distribution, one can determine, using equation (5), which consumers will demand connections and become subscribers. The market demand for calls is then obtained by summing the individual demand functions (equation (4)) over the set of subscribers.

To proceed, I specialize the general theory of consumer demand in the form of a particular model by postulating that utility is a quadratic function of the number of calls and a positive function of income,

$$(7) \quad U(x,q) = \lambda(y)\left[\nu + \alpha q - \frac{1}{2\beta}q^2\right],$$

$$\nu \geq 0, \alpha > 0, \beta > 0, \lambda'(y) > 0$$

which, for a given income level, results in a linear demand function

$$(8) \quad q = D(p) = (\alpha - p/\lambda)\beta$$

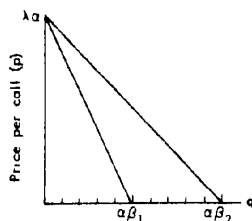



FIGURE 2 INDIVIDUAL DEMAND FUNCTIONS FOR CALLS OF TWO CONSUMERS WITH DIFFERENT TASTE PARAMETERS, BUT THE SAME INCOME y

and a reservation price function

$$(9) \quad R(p) = \lambda\left[\nu + \frac{1}{2}(\alpha - p/\lambda)^2\beta\right]$$

$$= \lambda[\nu + q^2/2\beta]$$

Although in many applications a quadratic utility function is unnecessarily restrictive, in this instance its satiation property is a virtue, since the demand for calls must be finite at a zero price. An alternative formulation, developed by Graham Pyatt, would be to include the value of time spent in telephone conversations in a generalized budget constraint.

When we normalize $\lambda = 1$ at an average level of income, the parameter ν represents the value of having telephone service when no outgoing calls are made, and thus the consumer's willingness to pay merely to receive calls or to have emergency communications available.[6] Under a flat rate tariff $L = L_o, p = 0$, the reservation price (which is the point of maximum possible utility) is $R_o = \lambda(\nu + \alpha^2\beta/2)$, and as shown in Figure 2 if $R_o \geq L_o$, the consumer subscribes and makes $q_o = \alpha\beta$ number of calls. Also note that, although we have represented λ as a function of income, in general, this parameter may depend on any number of exogenous factors that increase a consumer's willingness to subscribe to telephone service.

B. *The Effect of Income*

Higher levels of income lead to increased consumption of most goods and services,

[6]In this sense ν measures the consumer's option demand for telephone service

TABLE 1—PERCENTAGES OF HOUSEHOLDS WITH A TELEPHONE AVAILABLE, BY HOUSEHOLD INCOME AND PRICE OF BASIC SERVICE, 1970

	Price of Basic Monthly Service		
Household Income	Less than $5.00	$5.00 to $6.49	$6.50 and Above
Under $3,000	80.4	76.3	71.9
3,000–5,999	85.6	79.3	74.8
6,000–8,999	90.1	88.2	84.6
9,000–11,999	94.1	93.1	90.4
12,000 and above	97 9	96.8	96 9
All households	90.6	86 7	86.2

Source. See Perl, Table 12. *NERA* has adjusted the price of monthly service to estimate the minimum monthly rental available to each subscriber, see Perl, pp 20 22.

including telephone connections. A recent study by Lewis Perl matches data from the 1970 Census on telephone availability and household characteristics with price information from 100 revenue accounting office areas in the Bell System.[7] The percentage of households with a telephone available is cross tabulated with respect to the monthly charge for basic service, family income, age, sex, race, education, family size, and location. Perl's basic finding is that, irrespective of race, urbanism, or family type, telephone availability increases with both income and age, and decreases with higher monthly rental charges. Rural families, black families, and single males are less likely to have telephones than are urban families, families of other races, and married males. The effects of both income and price on the demand for telephone connections are apparent in Table 1. In my model, these empirical facts are incorporated in the assumption that $\lambda'(y) > 0$ and $\beta > 0$.

The effect of higher income on calling rates is not immediately apparent, since under a flat rate tariff, a subscriber can make any number of calls without changing his monthly bill, and thus the income he has available to spend on other goods. The limited available data on calling rates suggest that the number of calls per user *decreases* with higher income levels up to

[7]See Perl, filed by AT&T as Exhibit 21 with the Federal Communications Commission, 1974

TABLE 2—RESIDENTIAL TELEPHONE USE BY INCOME LEVEL

		Minutes per Day	
Household Income	Calls per Day per User	Originating (Local Only)	Terminating[a] (Local Plus Toll)
Under $3,000	3 48	6.18	8.24
3,000 5,000	2 76	4.72	5.75
5,000 8,000	1.45	4.91	6.30
8,000–10,000	1.60	5 06	5.41
10,000–15,000	1.42	5.28	5 64
15,000 20,000	1.52	4.84	4.72
20,000–30,000	0.97	4.30	4.85
Over $30,000	1.22	3.92	4.45

Source. AT&T, Subscriber Line Usage Study, May 1972 July 1973. Based on a sample of ten California exchanges using No 1 ESS switching equipment

[a]Incoming calls received.

about $15,000 per year (see Table 2). This initially surprising finding is consistent, however, with my assumption that a consumer's willingness to pay for telephone service increases with the number of calls he makes and with the data showing that fewer low-income than high-income households have a telephone. We may surmise that, for a *given* number of calls, the maximum monthly rate a low-income person is willing to pay for service is less than the rate a better-off person will pay. If so, low-income persons who do subscribe will be exactly those who make a disproportionate number of calls.

C. Market Demand

If all consumers had the same demands for telephone service, we could study any one of them to determine the aggregate effects of a tariff change. Consumer preferences are anything but equal, however. In traffic studies, telephone engineers find that the empirical frequency distribution of q_0, the number of residential calls per month under a flat rate tariff, follows a skewed, approximately lognormal distribution.

To obtain aggregate results, let us assume that the population consists of individuals who all have the same values of ν and α, but differing values of the taste parameter β, and differing incomes y. Equal values of α imply that all such consumers derive equal utility from the first call, and that the number of calls a subscriber makes is proportional to β. As a result, all subscribers at a given level of income have the same price elasticity of calls, $\eta_p^{q_i}$:

$$\eta_p^{q_i} = -\frac{p/\lambda}{\alpha - p/\lambda} = -\frac{p}{\lambda\alpha - p} \tag{10}$$

If we consider two consumers who have the same tastes β, but different incomes $y_1 < y_2$, we obtain demand curves like those in Figure 3. At any positive per call price, the demand of the lower-income consumer is more elastic and he makes fewer calls. At a zero price, however, both individuals are satiated at the same quantity.

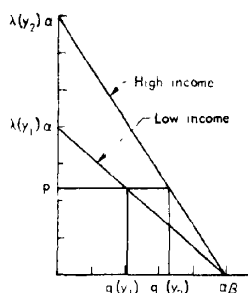



FIGURE 3. DEMAND FOR CALLS AS A FUNCTION OF INCOME, TWO CONSUMERS WITH IDENTICAL TASTES

Let $f_\theta(\beta)$ be the proportion of subscribers of type β. If we consider the consumer taste parameter β to be continuous, $f_\theta(\beta)$ is a statistical density function with parameters θ. Later, I will specify the form and parameters of this function to reflect the skewed distribution of calls that is empirically observed. For lack of any data to the contrary, I will further assume that this distribution is independent of the level of income.

Let us now derive the market demand curves implied by this model. First, let β_m be the value of β for the *marginal consumer*, whose surplus is just exhausted by the monthly charge $R(p, \beta_m) = L$, and who is therefore indifferent between telephone service and other goods. For this consumer,

$$\beta_m(L,p) = \frac{2(L - \lambda\nu)}{\lambda(\alpha - p/\lambda)^2} \tag{11}$$

and the number of calls made is

$$q_m = (\alpha - p/\lambda)\beta_m \tag{12}$$

In general, as income varies, the identity of the consumer at the margin, and therefore his tastes, will vary. A low-income consumer will be indifferent between subscribing and going without service only if he would make relatively extensive use of the telephone, whereas a high-income consumer will be willing to subscribe at a lower volume of calling. Since we have assumed that the underlying tastes for telephone service follow the same distribution regardless of income, it follows that (at positive

monthly rental rates L) a greater proportion of high-income than low-income persons will subscribe. It is also the case that the average number of calls made by the low-income consumers who do in fact subscribe will, for many tariffs, be greater than the average for high-income subscribers.

At each income level, a certain percentage of the consumers at that level will subscribe. The market demand for connections is the aggregation of these subscribers over all income levels, or

$$(13)\quad N(L,p) = M\int_y \int_{\beta_m(L,p)}^{\infty} f_\theta(\beta)\, d\beta \cdot g(y)\, dy = M\int_y [1 - F(\beta_m(L,p);\theta)] \cdot g(y)\, dy$$

where M is the total population, $g(y)$ is the probability density of consumers at each income level, and $F(\cdot)$ is the cumulative distribution function of β. For a given level of income, the demand-for-connection function $N(L,p)$ is of the form shown in Figure 4.

By differentiating equation (13) we can express the slopes of the connection-demand function at a given income level in terms of the parameters of the marginal consumer.[8] The slope with respect to the monthly charge is

$$(14)\quad \frac{\partial N}{\partial L} = -f_\theta(\beta_m)\frac{\partial \beta_m}{\partial L} = -\frac{f_\theta(\beta_m)\beta_m}{L - \lambda\nu}$$

and with respect to the per call charge is

$$(15)\quad \frac{\partial N}{\partial p} = \frac{-2f_\theta(\beta_m)q_m}{\lambda(\alpha - p/\lambda)^2} = q_m\frac{\partial N}{\partial L}$$

These slopes may be more conveniently related to each other in terms of the elasticities of demand:

$$(16)\quad \eta_p^N = \frac{E_m}{L}\eta_l^N$$

where η_p^N is the elasticity of demand for connection with respect to the per call price p, $E_m = pq_m$ is the variable expenditure on telephone calls by a marginal subscriber, and η_l^N is the elasticity of demand for connection with respect to the fixed charge L.

In a similar fashion, we may obtain the market demand for calls Q by aggregating the individual demands (equation (4)) of those consumers for whom $R(L,p) \geq L$. The market demand function is therefore

$$(17)\quad Q(L,p) = \int_y \int_{\beta_m(L,p)}^{\infty} q(p,\beta) f_\theta(\beta)\, d\beta \cdot g(y)\, dy$$

where $q(p,\beta) = [\alpha - p/\lambda(y)]\beta$ from equation (8). By differentiating equation (17) with respect to p, we obtain the slope of the aggregate demand curve at a given income level,

FIGURE 4. MARKET DEMAND CURVE FOR CONNECTIONS

[8] For these derivations we make repeated use of the derivatives

$$\frac{\partial \beta_m}{\partial L} = \frac{2}{\lambda(\alpha - p/\lambda)^2} = \frac{\beta_m}{L - \lambda\nu}$$

and

$$\frac{\partial \beta_m}{\partial p} = \frac{4(L - \lambda\nu)}{\lambda^2(\alpha - p/\lambda)^3} = \frac{2\beta_m}{\lambda(\alpha - p/\lambda)} = \frac{2q_m}{\lambda(\alpha - p/\lambda)^2}$$

$$(18)\quad \frac{\partial Q}{\partial p} = -\frac{1}{\lambda}\int_{\beta_m}^{\infty} \beta f_\theta(\beta)\,d\beta - (\alpha - p/\lambda)\beta_m f_\theta(\beta_m)\frac{\partial \beta_m}{\partial p}$$

and hence the elasticity of the aggregate curve at the specified level of income is

$$(19)\quad \eta_p^Q = \frac{p}{Q}\frac{\partial Q}{\partial p} = \frac{-p/\lambda \int \beta f_\theta(\beta)\,d\beta}{(\alpha - p/\lambda)\int \beta f_\theta(\beta)\,d\beta} - \frac{p}{Q}(\alpha - p/\lambda)\beta_m f_\theta(\beta_m)\frac{2q_m}{\lambda(\alpha - p/\lambda)^2}$$
$$= \frac{-p/\lambda}{\alpha - p/\lambda} + \frac{p}{Q}\frac{\partial N}{\partial p}(\alpha - p/\lambda)\beta_m$$
$$= \eta_p^{q_i} + \eta_p^N q_m/\overline{Q}$$

where $\eta_p^{q_i}$ is the price elasticity of calls for any current subscriber and $\overline{Q} = Q/N$ is the mean number of calls per subscriber. Thus, the elasticity of current subscribers is augmented by the effect of attracting new subscribers (in the case of a price decrease); the additional elasticity of the aggregate demand curve is proportional to the price elasticity of demand for connections, with the factor of proportionality expressing the calling rate of marginal subscribers relative to the average of all subscribers.

Finally, the derivative of the number of calls with respect to a change in the fixed monthly rate L is

$$(20)\quad \frac{\partial Q}{\partial L} = -(\alpha - p/\lambda)\beta_m f_\theta(\beta_m)\frac{\partial \beta_m}{\partial L} = \frac{-2\beta_m f_\theta(\beta_m)}{\lambda(\alpha - p/\lambda)}$$

and the elasticity is

$$(21)\quad \eta_L^Q = \frac{L}{Q}\frac{\partial Q}{\partial L}$$
$$= -\left(\frac{\beta_m f_\theta(\beta_m)}{L - \lambda\nu}\frac{L}{N}\right)\frac{2(L - \lambda\nu)}{\lambda(\alpha - p/\lambda)}\frac{1}{Q/N}$$
$$= \eta_L^N \frac{2(L - \lambda\nu)}{\lambda(\alpha - p/\lambda)^2}\frac{\alpha - p/\lambda}{\overline{Q}}$$
$$= \eta_L^N \beta_m \frac{q_m/\beta_m}{\overline{Q}} = \eta_L^N q_m/\overline{Q}$$

A reduction, say, in the monthly charge will increase the number of subscribers and thus increase the market demand for calls by the amount of their calls q_m per new subscriber. The percentage effect is the product of their number of calls relative to the average of current subscribers and the elasticity of demand for service.

To avoid encumbering the derivation with additional notation, we have obtained the aggregate relationships for a single level of income. The market demand curves and elasticities will consist of weighted averages of the slopes or elasticities at each income level in which the weights are simply the proportion of the total population at each income level.

D. *Restatement of the Demand Curves*

The relationships among the different elasticities of demand in the model, shown in equations (16), (19), and (21), are summarized as follows:

$$\eta_p^N = \frac{pq_m}{L}\eta_L^N$$

$$\eta_p^Q = \eta_p^{q_i} + \frac{q_m}{\overline{Q}}\eta_p^N$$

$$\eta_L^Q = \frac{q_m}{\overline{Q}}\eta_L^N = \frac{L}{pQ}\eta_p^N$$

Figure 4 shows the market demand curve for telephone connections as the flat rate tariff is varied. With no charge ($L = 0$), 100 percent of the households have telephones. At higher rates (above the monthly value of a telephone solely for incoming calls), market saturation N falls off. By using the frequency distribution of calls per month and the slopes of individual demand curves, we have calculated the proportion of consumers whose surplus at each price exceeds the monthly rental, and thus obtained the actual shape of the curve $N(L, 0)$.[9]

If a per call charge p_2 is added to the tariff, demand for connections is reduced to the left-hand curve $N(L, p_2)$ shown in Figure 4.

[9]Wherever possible, I denote a specific value of a price by use of a subscript (such as L_1), and the parametric value by the unsubscripted variable (L).

At any given level of the monthly charge, some former customers no longer find it worthwhile to subscribe. However, if the monthly fee were reduced to zero, everyone would have a telephone, even though there would be a charge for each call.

Let us now consider the market demand for calls Q. As just described, the level of the customer charge L determines the demand for telephone connections, and thus the number of customers. In doing so, it also affects the total number of calls made. Although a reduction in the flat rate tariff from L_1 to L_2 has no effect on any individual subscriber's rate of calling, Figure 5 reveals that it will induce a new group of consumers to subscribe and thereby increase the total number of calls made, from Q_1 to Q_2. Since, by assumption, the new subscribers value telephone service less highly and make fewer calls, the market demand for calls as a function of the monthly charge is less elastic than is the demand for connections.

The market demand curve for calls $Q(L,p)$, is of course also a function of the per call price p. Under a two-part tariff, the change in the total number of calls resulting from, say, a reduction in the per call rate will be the sum of two effects: (a) the aggregate increase in calling by present subscribers plus (b) the calls made by the new customers induced to subscribe as a result of the lower tariff. As a result, the market demand curve for calls is somewhat more elastic than is the demand curve of an existing subscriber.

FIGURE 5. MARKET DEMAND CURVE FOR CALLS AS A FUNCTION OF THE MONTHLY CHARGE

To appreciate the interrelationships of the demand curve for connections and the demand curve for calls, consider the effect of changing from a flat rate tariff $[L_1, 0]$ to a two-part tariff $[L_2, p_2]$ with a lower monthly rental charge $L_2 < L_1$. By itself, the reduction in L will increase both the number of subscribers and, by the amount of the new subscribers' calls, total calling as well. But the introduction of a per call charge lowers the attractiveness of service to new subscribers and reduces calling by all existing subscribers. The net effect is to move from the equilibrium N_1, Q_1 to N'_2, Q'_2 in Figures 4, 5, and 6.[10]

II. Benefits and Costs of Two-Part Telephone Tariffs

In this section I make an initial calculation of the gains and the additional costs of pricing local telephone service on a per call basis. To carry out a quantitative analysis, we require data sufficient to specify the parameters of both demand and supply functions for local calls as well as information on the additional costs of implementing a usage-sensitive tariff.

A. *Demand*

The literature contains few empirical studies of *U.S.* telephone demand in local markets. Because residential telephone service has almost always been marketed at flat rates, estimates of demand elasticities for calls are largely based on fragmentary

[10]The interdependence of the demand for calls and the demand for connections suggests that G. Franklin Mathewson and G. David Quirin's analysis of independent demand functions for connection and for calls is of limited applicability.

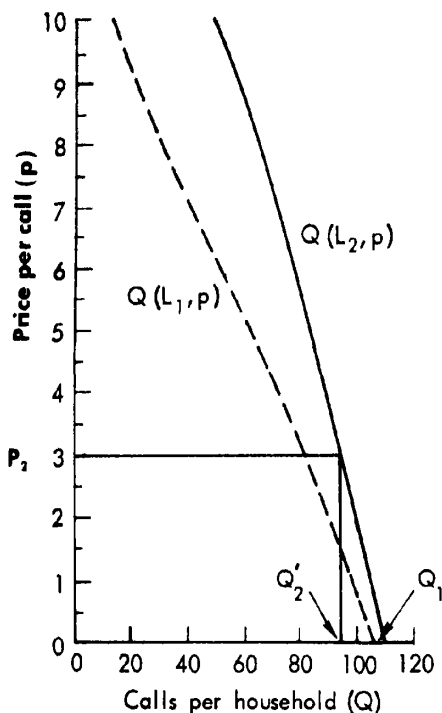



FIGURE 6. MARKET DEMAND CURVE FOR CALLS AS A FUNCTION OF THE PER CALL CHARGE

data from changes in public (coin) telephone tariffs. Robert Auray reports that price elasticity coefficients range from about −0.1 to −0.7 for Bell System data. In the United Kingdom, W. M. Turner's recent study reports local call price elasticities of less than 0.1 (in absolute value) for both residential and business subscribers. For long distance calls, the price elasticity of demand appears to be somewhat higher in both *U.K.* and *U.S.* studies. Using data from Illinois Bell, Littlechild and J. J. Rousseau calculate substantial own- and cross-price elasticities by time of day for state and interstate toll calls.

Cross-sectional and time-series variation in *U.S.* flat rates does permit the demand for telephone connections to be estimated directly. Table 1 above summarizes the cross-sectional evidence of the effects of price and income variations on telephone availability. These data imply that at the average level of flat rates, the price elasticity of demand for connections is about −0.1. By comparison, a time-series study of the demand for connections in the United Kingdom, where two-part tariffs are in effect, yields estimated price elasticities of −0.4 for residential service and −0.2 for business. The elasticities for the one-time hookup charge are estimated to be −0.2 for residence and −0.05 for business subscribers.

B. *Costs*

The marginal costs of supplying local telephone calls have been investigated in a quadratic programming model of the Illinois Bell Telephone Company (see Littlechild, 1970a; Littlechild and Rousseau). Because the provision of capacity to meet peak demand is a major part of total costs, marginal costs vary sharply by time of day. Littlechild and Rousseau found that in the short run, when capacity is fixed, marginal opportunity costs average 3¢ per call (5¢ day, 1¢ at night); in the long run, marginal costs average 2¢ per call (3¢ day, 1¢ at night).

The costs of connecting new subscribers to the telephone system (excluding their added costs of calling) derive principally from the additional main stations, subscriber lines, and dedicated central office equipment that is required. In Illinois Bell, these costs amounted to 36 percent of the total fixed and variable costs incurred in 1967, and overhead expenses accounted for an additional 10 percent. The costs of providing calling capacity and handling traffic amounted to 54 percent of the total.

Usage-sensitive pricing (*USP*) of local calls would require some additional fixed costs per subscriber as well as an increment to operating costs. The incremental capital cost required to measure local calling can range from as little as $5 to more than $50 per line, depending on the size of the local exchange, the type of switching equipment, and the extent of record keeping required. With mechanical switching equipment (i.e., step-by-step, panel, and cross-bar switches) additional hardware exhibits strong economies of scale, since there are substantial

fixed costs of equipment (for example, a minicomputer) and installation. In addition, the cost depends on whether calls are to be timed as well as counted, and whether records of the called and calling numbers must be retained for billing purposes. One manufacturer of metering equipment reports that the annualized hardware cost per line, when all lines are metered, decreases from more than $30 in a 1,000-line central office to just over $10 in a 10,000-line office. In small exchanges with only a few hundred subscribers, higher engineering and installation costs may bring the cost per line above $50. At the other end of the scale, New York Telephone estimates that the installation of metering equipment in its large No. 1 cross-bar exchanges in New York City has cost approximately $15 per line.

With an electronic central office, the cost of metering local calls is much lower, since electronic switching equipment is initially designed to handle this function. Once the computer programs have been written and debugged, the cost of metered service is relatively low. According to New York Telephone, the fixed conversion cost for an electronic central office ranges from $10,000 to $50,000—principally for software testing and debugging, initial purchase of computer tapes and setup charges. Since electronic switches today are installed only in large exchanges with several thousand lines, the cost per line for conversion to metered service should run no more than $2 to $5.

The additional operating costs for record keeping and billing will depend on whether full records of calling and called parties are kept. Estimates of these added operating costs vary widely, ranging between about $.001 and $.003 per call, or 10¢ to 40¢ per month for the average number of local calls. The added operating costs to time calls or to introduce peak and off-peak pricing would be quite small.

C. *Quantitative Analysis of Two-Part Tariffs*

I have calibrated the model as described in the Appendix to be consistent with the fragmentary empirical evidence on residential demand and cost conditions reviewed above. At a monthly flat rate of $5.61, the average rate in 1970, we assume that 87.3 percent of *U.S.* households subscribe to telephone service and make an average of 120 calls per month. On an overall per household basis, this corresponds to a mean of 105 calls per month. Table 3 lists the market demand functions generated by

TABLE 3—DEMAND FOR CONNECTIONS AND CALLS AVERAGE INCOME HOUSEHOLDS

Monthly Charge (L)	*A Demand for Connections* Percentage of Households Subscribing at Per Call Charge (p) of 0¢	2¢	4¢	6¢
$2 00	100.0%	99.9%	99.5%	98.2%
3 00	99.3	98.0	94 7	87 6
4 00	96 6	92 6	84.7	70 7
5 00	91 5	83 9	71 5	53 5
6.00	84.3	73 5	58 1	39 1
	B Demand for Calls Number of Monthly Calls per Household at Per Call Charge (p) of 0¢	2¢	4¢	6¢
$2 00	110 1	99.2	88.3	77.1
3 00	110.0	98.8	87.1	74.0
4 00	109.2	97.0	83.2	66.9
5 00	107 1	93.1	76.7	57 2
6 00	103 5	87.4	68.4	47.1

Note: Parameter values: $\nu = 0.5, \mu = 6.15, \sigma = 0.55, \alpha = 0.202, \lambda = 1.0$.

TABLE 4—ELASTICITIES OF DEMAND FOR CONNECTIONS
(Absolute Values)

Monthly Charge (L)	Elasticity When Per Call Charge (p) is: 0¢	2¢	4¢	6¢
A. η_L^N When Monthly Charge Varies				
$2.00	0.00	0.01	0.04	0.11
3.00	0.05	0.11	0.24	0.51
4.00	0.16	0.31	0.58	1.01
5.00	0.34	0.58	0.96	1.50
6.00	0.57	0.89	1.34	1.95
B. η_p^N: When Per Call Charge Varies				
$2.00	0	0.00	0.02	0.07
3.00	0	0.02	0.10	0.36
4.00	0	0.06	0.25	0.75
5.00	0	0.12	0.43	1.14
6.00	0	0.21	0.69	1.68

Note: Parameter values: See Table 3.

the model. These values were also used to plot Figures 4, 5, and 6. In each case, I have divided the total number of local calls by the number of households in the market, and reported the average number of calls on a *per household* basis rather than a per subscriber basis. Tables 4 and 5 show point elasticities of these demand functions, which vary with both the monthly charge and the per call charge.

Table 6 illustrates the effect of switching from a flat rate tariff at approximately the 1971 prevailing level ($5.50 per month) to one of two alternative two-part tariffs. (The two-part tariffs chosen for comparison provide approximately the same revenue to cover incurred costs as does the flat rate. The precise revenue requirement is imposed in the following section.) The change to either tariff leads to a sizeable increase in the proportion of average- and low-income households that subscribe to telephone service—an increase at average income levels from 88 percent to at least 98 percent.

TABLE 5—ELASTICITIES OF DEMAND FOR CALLS
(Absolute Values)

Monthly Charge (L)	Elasticity When Per Call Charge (p) is: 0¢	2¢	4¢	6¢
A. η_L^Q: When Monthly Charges Varies				
$2.00	0.00	0.00	0.01	0.03
3.00	0.01	0.03	0.08	0.21
4.00	0.05	0.12	0.25	0.53
5.00	0.13	0.26	0.50	0.89
6.00	0.25	0.45	0.77	1.25
B. η_p^Q: When Per Call Charge Varies				
$2.00	0	0.11	0.25	0.42
3.00	0	0.12	0.28	0.57
4.00	0	0.13	0.36	0.81
5.00	0	0.16	0.47	1.10
6.00	0	0.20	0.60	1.39

Note: Parameter values: See Table 3.

TABLE 6—CHANGES IN NUMBER OF SUBSCRIBERS, CALLING RATES, AND MONTHLY BILLS UNDER TWO-PART TARIFFS

Item	Flat Rate (L = \$5.50, p = 0¢)	Measured Rate: Alternative I (L = \$3.00, p = 2¢)	Measured Rate: Alternative II (L = \$2.00, p = 3¢)
Average-Income Households (λ = 1.0)			
Number of subscribers[a]	88 1	98.0	99.7
Calls per subscriber per month	120	101	94
For original subscribers:			
Calls per subscriber per month	120	108	102
Monthly bill	\$5.50	\$5.16	\$5.06
For new subscribers			
Calls per subscriber per month	-	38	33
Monthly bill		\$3.76	\$2.99
Low-Income Households (λ = 0 8)			
Number of subscribers[a]	77.0	93 1	98.2
Calls per subscriber per month	129	102	91
For original subscribers.			
Calls per subscriber per month	129	113	105
Monthly bill	\$5 50	\$5 26	\$5.15
For new subscribers			
Calls per subscriber per month		47	40
Monthly bill	–	\$3.94	\$3.20
High-Income Households (λ = 1 4)			
Number of subscribers[a]	97 0	99.8	100.0
Calls per subscriber per month	113	103	99
For original subscribers			
Calls per subscriber per month	113	105	101
Monthly bill	\$5 50	\$5.10	\$5 03
For new subscribers:			
Calls per subscriber per month	-	36	28
Monthly bill		\$3.72	\$2.84

Note Parameter values: See Table 3
[a] Percent of all households

At the same time, the imposition of either a 2¢ or 3¢ per call charge reduces the number of calls. For example, the 88 percent of all average-income households who now subscribe under a flat rate and make on average 120 calls per month, would make only 108 calls at 2¢ per call or 102 calls at 3¢. The new telephone subscribers who are attracted by the reduced monthly charges make considerably fewer calls: either 38 or 33 calls per month at average levels of income.

Table 6 also shows the effects on existing subscribers of switching to a two-part tariff. For households at all income levels, monthly bills decline somewhat from the flat \$5.50 rate. Low-income households are most sensitive to the imposition of a per call charge and make the greatest reduction in calling. However, because low-income subscribers under flat rates are also systematically high-volume users, they nevertheless make more calls under measured rates than do average- or higher-income subscribers, and as a result pay slightly higher bills (\$5.26 vs. \$5.16 or \$5.10 under the 2¢ per call rate). At the same time, an important effect of the two-part tariff is to make telephone service available to low-income households who could not previously afford it. As noted, these persons are predominantly low-volume users whose bills average only some \$3.20 to \$4.00 under measured service.

D. *Optimal Two-Part Tariffs*

Let us now consider the optimal *single tariff* to be offered to the residential class of customers.[11] It is assumed that total costs of supplying residential service can be represented as

$$C = F + sN + rQ \tag{22}$$

where F is fixed costs, s is the constant marginal cost per subscriber, and r is the constant marginal cost per call.

To maximize economic welfare—the sum of consumer's plus producer's surplus—the ideal two-part tariff would set each part of the tariff equal to the respective marginal cost of a connection and a call:

$$L^* = s, p^* = r \tag{23}$$

Under this tariff, the revenue obtained from each subscriber, $L^* + p^*q_i$, will just cover his marginal cost $s + rq_i$ to the firm. Furthermore, each customer will be induced to make telephone calls until the value to him of the last call is equal to the incremental cost r of supplying another call. Every consumer who is willing to pay the incremental cost of telephone service is able to purchase it. Using this marginal-cost tariff the telephone company will incur a deficit equal to its fixed costs F.

When the tariff must fully recover costs, the second best optimal two-part tariff will necessarily involve pricing at least one component of telephone service above its marginal cost. In this case, because of the nonzero elasticity of demand for connections, too high a monthly charge L will keep some potential subscribers out of the market who are nevertheless willing to pay at least the marginal cost of their service, $s + rq_i$. But by setting p greater than r, the telephone company can increase its revenue from per call charges and therefore reduce the monthly charge to attract more subscribers. In general, the optimal two-part telephone tariff that satisfies the revenue constraint will have a higher per call charge than would be indicated by pure marginal cost ($p = r$) pricing of calls.[12]

To calculate the optimal two-part tariff, I will limit my analysis to residential consumers and assume that any alternative tariff must be constrained to raise the same revenue (except for differences in costs due to changes in the number of subscribers and calls) that is now raised by flat rates from residential subscribers as a group. This procedure has the advantage of making alternative tariffs directly comparable with current pricing practices, but it is likely to result in a conservative estimate of the welfare gains that could be achieved by simultaneously optimizing both residential and business tariffs for local service. Studies of *U.S.* local exchange costs have generally found revenues from residential customers insufficient to cover their costs of service (see, for example, Walter Baer and the author), and it is likely that a higher residential revenue constraint would increase the overall efficiency with which telephone resources are used.

To examine the sensitivity of the magnitude of possible welfare gains from two-part tariffs to the specification of the model, I have made a series of alternative assumptions about cost and demand parameters. It is assumed that marginal costs are either 2¢ or 3¢ per call, and that the marginal costs per additional subscriber are either 60 or 80 percent of the remaining revenue requirement ($5.61 per subscriber at average levels of usage for flat rate service); the remaining costs are fixed. Three combinations of these assumptions provide us with

[11] Except for the paper by Gerald Faulhaber and John Panzar, the additional welfare gains that may be achieved by allowing the consumer to select among a set of tariffs have as yet received little theoretical analysis. Optional tariffs have considerable importance for policy. For example, it is frequently proposed that a "lifeline" tariff, with a low monthly rental charge but a high per call price, be offered as an alternative to flat rate service.

[12] If it happened that marginal subscribers and existing subscribers made the same number of calls, on average, then $p = r$ would be optimal (See Yew-Kwang Ng and Mendel Weisser.) This could occur only if the demand curves for different consumers were to intersect.

the following cost parameters:

Fixed Cost per Household:	Marginal Cost per Subscriber	Marginal Cost per Call
$1 12	$1.92	2¢
0 35	1.60	3
0 70	1 20	3

Fixed costs are stated on a *per household* basis, since they are invariant to the number of actual subscribers. Similarly, we will also measure welfare per household in order to make direct comparisons between tariffs with differing numbers of subscribers. The aggregate level of fixed costs or welfare would be obtained by simply scaling up those values by the number of households in the local exchange area.

Using the parameters of the demand function shown in Table 3 and in the Figures 4, 5, and 6, we calculate the aggregate welfare—consumer's plus producer's surplus produced by the $5.61 flat rate tariff representative of *U.S.* residential tariffs at 1971 price levels. To make this calculation, we add up the welfare gains at each level of calling weighted by the number of subscribers making that many calls, and divide by the number of households. We thus obtain a monthly value of $6.16 per household for the average welfare of telephone service supplied under flat rates. This value, as well as the demand for connections and calls, are shown in the first row of Table 7.

In contrast to this baseline case, two-part tariffs result in more subscribers, reduced calling, and net gains in welfare. The remainder of Table 7 reports the optimal tariffs, market demands, and welfare changes under a variety of alternative cost and demand assumptions. In each case, the optimal tariff is calculated by maximizing aggregate welfare, measured at the average level of income, subject to the constraint that revenues cover costs. (The algorithm is outlined in the Appendix.) The second, third, and fourth rows of Table 7 show the optimal tariffs under each of the variations in cost assumptions. For example, with a marginal cost of 2¢ per call, the optimal tariff is a monthly charge L = $2.71 and a per call charge p = 2.34¢. The increment in total welfare, 5.2 percent or $0.32 per

TABLE 7 OPTIMAL TARIFFS UNDER DIFFERENT COST CONDITIONS

Marginal Cost			Optimal Tariff		Market Demand		Welfare Measure	
Per Subscriber (s)	Per Call (r)	Metering Cost	Monthly Charge (L)	Per Call Charge (p)	Subscribers, Percent of Market (N)	Number of Calls per Household per Month (Q)	Per Household (W)	Percent Change from Flat Rate ($\Delta W/W_0$)
Baseline case			$5 61	flat rate	87.3	105 1	$6 16	
$1 92	2 0¢	none	2.71	2.34¢	98 7	97 1	6 48	5.2
1.60	3 0	none	1 90	3.04	99 8	93 5	6 62	7 5
1 20	3 0	none	1 81	3 08	99 9	93 3	6 67	8.3
1 92	2 0	low	2 75	2 46	98 5	96 4	6.33	2 8
1 60	3 0	low	1 95	3 14	99 7	93 0	6.47	5 2
1 20	3 0	low	1 87	3 18	99 8	92 8	6 53	6.0
1 92	2 0	high	2 79	2 70	98 1	95 0	6.06	−1.6
1 60	3 0	high	2 03	3 34	99 6	91 9	6 21	0 9
1 20	3 0	high	1 92	3 40	99 7	91 6	6.26	1.7

Notes Demand Parameters

α	0 202
ν	0 5
μ	6 15
σ	0.55

Metering Costs

Level	Per Subscriber per Month	Per Call
none	0	0
low	5 0¢	0.1¢
high	12.5¢	0.3¢

household per month, is the result of eliminating calls valued at less than their marginal cost and gaining new subscribers whose consumer's surplus is positive. Note that, as discussed above, the 2.34¢ per call charge in the optimal tariff is greater than the 2.0¢ marginal cost of a call. (With a "marginal cost" tariff (L = $3.06, p = 2.0¢), the welfare gain is about 2 percent lower.) Under the other cost assumptions (third and fourth rows), welfare gains from two-part pricing range up to 8.3 percent.

However, such gains must be compared with the added costs of metering and billing for each call. If the costs of new metering and billing services are at the low end of the range discussed earlier ($0.05 monthly per subscriber plus 0.1¢ per call) and marginal costs per call are 2¢, the optimal tariff, shown in the fifth row, increases to a monthly charge of $2.75 per month plus 2.46¢ per call. At these rates, slightly fewer households will subscribe to service, and there will be fewer calls. Welfare would be $6.33 per household per month, a gain of 2.8 percent over the flat rate situation. But if the added capital and operating expenses are considerably larger ($0.125 per subscriber plus 0.3¢ per call), then, as shown in row 8 of the table, added costs can more than offset the efficiency gains of the two-part tariff, and a 1.6 percent loss of welfare results.

To test the sensitivity of the results, I have made similar calculations for three other sets of values for demand parameters besides those used in the initial case.[13] Examination of all of these cases reveals that the welfare gains from two-part tariffs fall in the following intervals:

Metering Cost	Range of Welfare Changes
None	4% to 13%
Low	-1% to 9%
High	-3% to 3%

In general, two-part tariffs lead to the greatest improvement in economic efficiency when metering and implementation costs are low and the marginal calling costs are relatively high (3¢ rather than 2¢ per call). At the upper end of the feasible range, a 9 percent welfare gain is equivalent to an annual gain in economic efficiency of about $5 per household. If two-part tariffs were eventually applied to residential service in all metropolitan exchanges, the nationwide gains would be on the order of $250 million per year.

For several reasons, the calculations reported here are likely to understate the full magnitude of the welfare gains that can be achieved by two-part tariffs. First, although my model assumes that consumer's individual demand curves are linear (Figure 1), it is likely that for many consumers the marginal utility of additional calls (or minutes of conversation) declines at a more than linear rate and results in a convex demand curve. In this case, metered service will lead to a more pronounced reduction in calling and greater efficiency gains than predicted by the linear demand curve model. Second, as noted earlier, we have restricted our analysis to tariffs that maximize welfare while maintaining the historical pattern of revenue raised from the residential subscriber class. Simultaneously optimizing local service tariffs for both residential and business users would further increase the efficiency of the telephone system, although it would carry with it the possibility of higher residential telephone bills.

Finally, I have made no attempt to measure the additional efficiency gains that could be realized by incorporating peak-load pricing into measured local service tariffs. Since peak load tariffs would impose almost no further costs in the large electronic exchanges once the equipment for metering two-part tariffs is available, any shifts in calling patterns that reduced the maximum traffic-handling capacity needed would represent additional net gains in welfare In Norway, daytime (8 A.M.–5 P.M.) peak load rates for minutes of local calls went into effect in January 1975; since then, residential and business users have decreased the number of their peak-period calls by 17 and 10 percent, respectively, and

[13]These results may be found in the author, Table 7, p. 34.

off-peak use has increased about 8 percent. In addition, residential and business users have shortened their peak-hour calls by some 40 percent and 7 percent, respectively (see Jan-Erik Kosberg, Olav Gaustad, and Kristian Bø). This degree of demand elasticity, combined with the ranges of peak and off-peak marginal cost estimates described earlier, suggests that peak load pricing applied to residential customers in metropolitan areas would double or triple the welfare gains achieved from a two-part tariff alone. Thus, considering only residential customers, the potential welfare gains from usage-sensitive pricing are likely to exceed $500 million per year

III. Conclusions

The available evidence on demand and cost conditions for local telephone service in the United States suggests that per call charges for local telephone calls will increase social welfare in the exchanges, primarily those with electronic switches, that have low metering costs In magnitude, the gains calculated from our model are rather modest, but they should be considered minimum estimates of what can be achieved with measured service for all customers and with the opportunity to offer discount rates for calling outside of the peak load period.

The effect of measured service on different subscribers will depend on their patterns of use. In general, measured service will attract new subscribers and allow the telphone company to achieve virtually universal service. All groups of consumers will make some reductions in their calling. Persons who make substantially more than the average number of calls will pay higher bills, while low-volume callers will save under two-part tariffs. These patterns will apply at both high and low levels of income, but generally speaking, current low-income subscribers who tend to use their telephones more will make greater reductions in their calling, and will on average pay about the same amount for service as higher-income subscribers. At the same time, new subscribers will be predominantly those lower-income households who make relatively few calls and who have therefore been unwilling to subscribe to telephone service because they have regarded it as too expensive under current flat rates.

In some respects, this paper has raised more questions than it has answered with regard to the pricing of local telephone service. For example, what welfare gains would peak load pricing offer? In what circumstances are multiple tariffs, which permit the consumer to choose, for example, between flat-rate and measured service, superior to a single two-part tariff? What biases does rate of return regulation inject into the pricing decisions of telephone utilities? Each question calls for new theoretical inquiries.

With respect to empirical research, perhaps the greatest need is for econometric studies of telephone demand. Promising possibilities would include examination of exchanges that have implemented usage-sensitive pricing for residential customers, and statistical analysis of the distribution of calls and holding times in the context of a behavioral model. Data from such studies would permit the model developed in this report to be extended to incorporate peak and off-peak calling periods. Finally, new insights for *U.S.* practice could be gained from a comparative analysis of the telephone pricing and demand experience of foreign utilities that have long operated with various forms of measured service.

Appendix

In this Appendix, the major computational steps required to calibrate the model and calculate optimal tariffs are outlined. Further details are available in my 1976 paper.

A. *Calibrating the Model*

On the basis of statistical studies of telephone traffic patterns by Hans Kraepelien and Bell System personnel, it is assumed that $f_\theta(\beta)$ is lognormal with unknown $\theta = (\mu, \sigma^2)$ for the mean and variance. Krae-

pelien's work suggests that σ lies roughly in the 0.3 to 0.6 range. Given a value of σ, knowledge of the arithmetic mean number of calls by all potential customers determines the parameter μ. We use a value of 120 calls per actual subscriber per month for customers receiving flat rate service in metropolitan areas.

In 1970 the U.S. Census reports that 87.3 percent of residential households had a telephone available, and Perl reports that in 1971 the average monthly charge for basic residential service in Bell System revenue districts was \$5.61.

Using these values,

$$N_o = 0.873$$
$$\bar{Q}_o = Q_o/N_o = 120$$
$$L_o = 5.61$$

we employ the following algorithm to determine the parameters ν, α, μ, and σ^2 at an average level of income ($\lambda = 1$):

1) Set initial values of ν and σ (Runs were made using $\nu = 0$ to 2.0 and $\sigma = 0.3$ to 0.6.)

2) Select a trial value of μ and determine the value of the lognormal variate for the marginal consumer

3) Calculate the mean number of calls per household, making use of the distribution theory for a truncated lognormal variable.

4) Compare the calculated and actual number of calls per subscriber and iterate for different values of μ until the discrepancy is small.

At the conclusion of step 4) we have obtained one set of parameter values (ν, σ, μ, α) that is consistent with the observed quantities of connections and calls demanded at the \$5.61 flat rate. Now evaluate the demand functions for connections and for calls, and their point elasticities, for an entire set of flat rate and two-part tariffs. The one further piece of market data available suggests that the elasticity of demand for connections with respect to the monthly charge is in the range of -0.1 to -0.4. If the calculated demand functions are in agreement with this range of elasticities at the initial flat rate, we have determined one set of calibrated parameter values.

By returning to step 1) and selecting other starting values for ν and σ, we then obtain several sets of model parameters that are broadly consistent with the available empirical facts.

To calibrate the parameter λ of the model, data (see Table 1) are used from Perl's study of telephone availability to determine values of the function $\lambda(y)$ at selected levels of income. Under flat rate tariffs in the \$5.00 to \$6.49 range, Perl finds that availability ranges from 75 percent at incomes below \$3,000 to 97 percent for incomes above \$12,000. We chose the following availability rates to calculate representative values of λ for low- and high-income consumers:

Income	Availability
Low	75.5%
Average	87.3
High	96.7

To obtain the required values of λ, we first find the values of the taste parameter β_m for the marginal consumer that result in each of these levels of availability and then solve for the corresponding λ under the flat rate tariff. Final computer runs were made for the four sets of parameter values summarized in Appendix Table 1.

B. *Calculation of Optimal Tariffs*

We calculate the welfare-maximizing tariff for a given set of cost and demand

APPENDIX TABLE 1

Parameter	Case I	II	III	IV
σ	0.55	0.40	0.50	0.60
μ	6.15	6.40	6.20	6.10
α	0.202	0.172	0.201	0.202
ν	0.5	0	0	1.0
λ				
low income	0.80	0.83	0.80	0.80
average income	1.00	1.00	1.00	1.00
high income	1.41	1.32	1.41	1.41

parameters by (a) determining each two-part tariff that satisfies the constraint that revenues equal costs, (b) evaluating the aggregate consumer's plus producer's surplus of telephone subscribers under that tariff, and then (c) choosing the tariff that gives rise to the greatest surplus.

Total cost is made up of fixed cost F, a constant marginal cost per subscriber s, and a constant marginal cost per call r: $C = F + sN + rQ$. Total revenue is $R = LN + pQ$. The market demand quantities the number of subscribers N and number of calls Q are determined by evaluating the demand functions $N(L,p)$ and $Q(L,p)$ in equations (13) and (17) at the average income level ($\lambda = 1$). For convenience, N is expressed in terms of the percentage of all households who subscribe, and, in order to have a measure that is invariant to the number of actual subscribers, we measure Q as the number of calls per household

For tariffs satisfying the budget constraint $R = C$, aggregate surplus will be the sum of the areas under the individual demand curves of actual subscribers less the total costs of providing service. For average income households ($\lambda = 1$) this is

$$W = \nu N + \frac{1}{2}(\alpha + p)Q - C$$

To calculate the optimal tariff we use the following algorithm:

1) Set the values of the cost parameters (F,s,r) and demand parameters $(\nu, \sigma, \mu, \alpha)$.

2) Set initial tariff parameters $L = L_o = 5.61$, $p = 0$, verify that $R = C$, and calculate W

3) Reset s and r to include metering costs. Set $p = r$ and set a trial value of $L = L_o$.

3.1) Calculate R and C, and if $R > C$, iterate for lower value of L.

3.2) When $|R - C| < \epsilon = \$0.01$, a feasible tariff $[L,p]$ has been found. Calculate W.

3.3) Replace p by $p + \Delta p$, reset $L = L_o$, and return to step 3.1.

4) Find the maximum W over the set of feasible tariffs.

In the simulations reported here, two levels of marginal costs per call are used: $r = 2¢$ and $r = 3¢$. If marginal calling costs are rQ_0 at the initial level of usage under flat rate service, then we assume that the remaining costs to be recovered are equal to total revenues R less these costs, or $Z = R - rQ_0$. These costs are then distributed among fixed costs (F) and per subscriber costs (s) by assuming that either 20 or 40 percent of such costs are entirely fixed costs, and the balance are marginal costs per subscriber; these percentages were chosen to correspond roughly to the proportions reported from Illinois Bell by Littlechild and Rousseau

REFERENCES

John Aitcheson and James A. C. Brown, *The Lognormal Distribution*, Cambridge 1957.

R. Artle and C. Averous, "The Telephone System as a Public Good: Static and Dynamic Aspects," *Bell J. Econ.*, Spring 1973, *4*, 89–100

R. R. Auray, "Elasticity of Demand for Communications Services," in *Economics of the Regulated Communications Industry in the Age of Innovation*, Boston 1970.

W. S. Baer and B. M. Mitchell, "The Economic Impact of Competition on an Independent Telephone Company," *Publ. Util. Fortnightly*, Oct. 23, 1975, *96*, 1–7.

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83

P. J. Berman and A. G. Oettinger, "The Medium and the Telephone: The Politics of Information Resources," work. paper no. 75-8, Program Inform. Techn. Publ. Pol., Harvary Univ. 1975, 30a.

S. K. Black and P. V. Tryon, "Increased Telephone Installation Rates: A Statistical Analysis of Colorado's 1972 Rate Change," rept. no. 76-90, Office of Telecommunications, U.S. Department of Commerce, Washington, June 1976.

John Brooks, *Telephone: The First Hundred Years*, New York 1976.

L. Carlton, "Sensitizing Telephone Rates,"

Telephony, No. 11, 15, Sept 1975, *189*, 66 70.

G. R. Faulhaber and J. C. Panzar, "Optimal Two-Part Tariffs with Self-Selection," economic disc paper no. 74, Bell Labs, Jan. 1977.

M. S. Feldstein, "Equity and Efficiency in Public Sector Pricing. The Optimal Two-Part Tariff," *Quart. J. Econ.*, May 1972, *86*, 175 87.

L. Garfinkel, "Usage Sensitive Pricing: Studies of a New Trend," *Telephony*, No. 66, Feb. 10, 1975, *188*, 24 29.

H. S. E. Gravelle, "Telephone Rentals: Comment," *Appl Econ.*, 1972, *4*, 235 38

A. Hazlewood, "Optimum Pricing as Applied to the Telephone Service," *Rev Econ. Stud.*, No. 2, 1951, *18*, 67 78.

J.-E. Kosberg, O. Gaustad, and K. Bø, "Some Traffic Characteristics of Subscriber Categories and the Influence from Tariff Changes," paper presented at the Eighth International Teletraffic Congress, Melbourne, Australia, Nov. 1976.

H. Y. Kraepelien, "The Influence of Telephone Rates on Local Traffic," paper presented at the Second International Teletraffic Congress, The Hague, July 1958.

S. C. Littlechild, (1970a) "Peak-Load Pricing of Telephone Calls," *Bell J Econ.*, Autumn 1970, *1*, 171 210.

——, (1970b) "A Game-Theoretic Approach to Public Utility Pricing," *Western Econ. J.*, June 1970, *8*, 162 66.

——, (1970c) "A Note on Telephone Rentals," *Appl Econ.*, 1970, *2*, 73-74.

——, "Two-Part Tariffs and Consumption Externalities," *Bell J. Econ.*, Autumn 1975, *6*, 661 70.

—— **and J. J. Rousseau,** "Pricing Policy of a U.S. Telephone Company," *J. Publ Econ.*, Feb. 1975, *4*, 35 56.

M. G. Marchand, "The Economic Principles of Telephone Rates Under a Budgetary Constraint," *Rev Econ. Stud.*, Oct. 1973, *40*, 507 15.

G. F. Mathewson and G. D. Quirin, "Metering Costs and Marginal Cost Pricing in Public Utilities," *Bell J. Econ.*, Spring 1972, *3*, 335-39.

D. J. Mayston, "Optimal Licensing in Public Sector Tariff Structures," in Michael Parkin and A. R. Nobay, eds., *Issues in Contemporary Economics*, Manchester 1974.

B. M. Mitchell, "Optimal Pricing of Local Telephone Service," Rand Corp., R-1962-MF, Nov. 1976.

Y.-K. Ng and M. Weisser, "Optimal Pricing with a Budget Constraint—The Case of the Two-Part Tariff," *Rev. Econ. Stud.*, July 1974, *41*, 337-45.

L. J. Perl, "Economic and Demographic Determinants of FCC Telephone Availability," National Economic Research Associates (*NERA*), Inc., April 5, 1975, filed by AT&T Company, FCC Docket No. 20003, Bell Exhibit 21.

T. Pousette, "The Demand for Telephones and Telephone Services in Sweden," paper presented at the European Meetings of the Econometric Society, Helsinki, Aug. 1976.

G. Pyatt, "Some Economics of a Public Utility," disc. paper no. 28, Univ. Warwick, Sept. 1972

J. Rohlfs, "A Theory of Interdependent Demand for a Communications Service," *Bell J Econ*, Spring 1974, *5*, 16 37.

W. G. Shepard, "Residence Expansion in the British Telephone Service," *J Ind. Econ.*, July 1966, *14*, 263-74.

L. Squire, "Some Aspects of Optimal Pricing for Telecommunications," *Bell J. Econ.*, Autumn 1973, *4*, 515 25.

W. M. Turner, *CEPT Study of the Growth of the Telephone Service, Report by United Kingdom*, Telecommunications Headquarters Marketing Dept., British Post Office, London 1975.

B. Von Rabenau and K. Stahl, "Dynamic Aspects of Public Goods: a Further Analysis of the Telephone System," *Bell J. Econ.*, Autumn 1974, *5*, 651-69.

American Telephone and Telegraph Company, *Annual Statistical Report for 1974*, New York 1975.

——, *Subscriber Line Usage Study, May 1972-July 1973*, New York 1973.

Federal Communications Commission, *Notice of Inquiry in Docket No. 20003*, 46 FCC 2d

214, Apr. 1974.

National Association of Regulatory Utility Commissioners, *Exchange Service Telephone Rates in Effect June 30, 1974*, Washington 1974.

Seattle Post-Intelligencer, "New AT&T Plan on Phone Rates," Nov. 30, 1976, pp. 1, 10.

Washington Star News, Dec. 4, 1973, p. A-12.

Endogenous Bias in Technical Progress and Environmental Policy

By ROGER A. MCCAIN*

Environmental economics is based on an analysis of external social costs in an essentially neoclassical context, with a given range of techniques some of which imply the destruction of environmental amenities. The set of available techniques may be dense, so that smooth substitution of capital, labor, and environmental degradation is possible; or it may not be dense,[1] but in any case it is given. The neoclassical theory of economic growth, by contrast, largely rests on the assumption of improving technology (see Robert Solow). This dichotomy is remarkable in that many laypersons and some economists believe that economic growth and environmental degradation are associated. One of the most important insights of the received theory of environmental economics is, indeed, that pollution would occur in an unregulated stationary economy. However, that should not close the issue insofar as environmental policy is concerned. It is possible that technical progress may be biased in the direction of environmental destruction, as it may be biased in the direction of saving labor, of saving capital, etc. (see John R Hicks). Indeed, some literature in the theory of economic growth assumes that such bias is endogenous, determined by market forces.[2] If so, then environmental policy will have some impact on the bias, and this should be taken into account in the consideration of environmental policy.

This paper considers two approaches to endogenous bias in technical progress with a view to their implications for environmental policy. The first section explores the innovation-possibility frontier approach, and the second explores a newer approach grounded in criticism of the innovation possibility frontier approach. The third section summarizes and restates the conclusions.

I

Economists have speculated for some time that the characteristics of new technology respond to the forces of the market (see Hicks; Fellner, 1961; Jacob Schmookler). This speculation does, however, raise troublesome conceptual issues. To say that new techniques respond to market forces is to say that the new techniques are chosen to be suitable to the market forces. Yet, clearly, the choice cannot be perfectly free. There must be some sort of limitation on the set of available *technological opportunities*. The *innovation-possibility frontier* is an idealization of such a constraint.

We consider the existing technology as being characterized by a production function of the form

$$Q = f(EN, FK) \tag{1}$$

with constant returns to scale and a constant elasticity of substitution. In particular, since we shall have to consider the case of a factor of production which has a zero price, we assume

$$Q = \min(EN, FK) \tag{1'}$$

so that the elasticity of substitution is zero. Technical progress takes the form of changes in E and F, constrained by the innovation-possibility frontier:

$$g(E^*, F^*) = 0 \tag{2}$$

*Associate professor of economics, Temple University. Preparation of an earlier version was supported by the Faculty Senate Fund of the City College of the City University of New York. I am indebted to an anonymous referee whose suggestions improved the paper.

[1] An input-output system with countably many techniques of production, for example, is not dense.

[2] See Charles Kennedy, William Fellner (1961, 1967), also Syed Ahmad, Gary Becker, the author (1970, 1974a), William Nordhaus, and Horst Reichenbach.

where E^* denotes $(1/E)(dE/dt)$, F^* denotes $(1/F)(dF/dt)$, and g is a convex function.[3] The characteristics of this model are now well known (see Kennedy). The key conclusions are 1) that for an optimum,

$$\frac{wN}{rK} = (\partial g/\partial E^*)/(\partial g/\partial F^*) \tag{3}$$

where w is the wage and r the rate of return; and 2) in steady-growth "equilibrium,"

$$F^* = 0 \tag{4}$$

that is, all technical progress is labor augmenting.

The steady-growth equilibrium requires that equation (4) be fulfilled, if the average propensity to save is constant. The constant propensity to save implies that Q/K is constant, and so,

$$Q^* = K^* \tag{5}$$

However, we may assume that the constraint

$$Q = FK \tag{6}$$

is always binding, since otherwise a small F^* and correspondingly larger E^* would have been optimal in some previous period. From (6) follows

$$Q^* = F^* + K^* \tag{7}$$

from which (4) follows as a property of steady growth.

It seems plausible that the underpricing of environmental amenities would result in a process of technical change biased in the direction of environmental pollution. To explore this point, we might simply add environmental capacity,[4] an unpriced collective good, as a third input in (1) and (2) above, yielding

$$Q = f(EN, FK; GR) = \min(EN, FK, GR) \tag{8}$$

$$g(E^*, F^*, G^*) = 0 \tag{9}$$

with R interpreted negatively as environmental capacity used up, or positively as the quantity of industrial pollution. With environmental capacity priced at zero, and with $F^* = 0$ in a steady-growth process, we might expect to find a situation such as that portrayed in Figure 1. There, ff' is the cross section of the innovation-possibility frontier with $F^* = 0$. If environmental amenity is unpriced, a profit-maximizing entrepreneur (or a bonus-maximizing socialist manager) will maximize E^*, yielding equilibrium rates of technical progress G_0^*, E_0^*, with $G_0^* < 0$ a possibility. We then have, from $Q = GR$,

$$R^* - Q^* = -G_0^* \tag{10}$$

that is, pollution outgrows output by the absolute value of the rate of pollution augmentation G_0^*.

If government imposes a price on pollution of u per unit, pollution cost of uR must

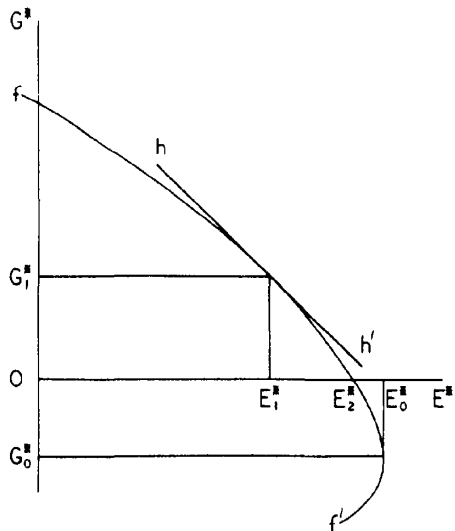



FIGURE 1

[3] As a notational convention, for any variable x, $(1/x)(dx/dt)$ is denoted x^*.

[4] Herman Daly has stressed that the notions of ultimate input and final output are deceptive. The level of production is a level of throughput, and this throughput is constrained, in the long run, by the capacity of the environment to absorb and recycle wastes. This is the rationale of the term "environmental capacity" and the treatment of it as in some sense fixed.

be balanced against the cost of labor wN. The slope of line hh' is wN/uR; and, subject to that expenditure ratio, the profit-maximizing E^*, G^* pair is E_1^*, G_1^* with $G_1^* > 0$. This seems to offer the possibility of stabilizing pollution R in a growing economy. What is required is that u be large enough (the slope of hh' shallow enough) that

$$G_1^* = Q^* = E_1^* + N^* \tag{11}$$

for identically

$$G^* + R^* = Q^* \tag{12}$$

so it would follow that[5]

$$R^* = 0 \tag{13}$$

In fact, however, a steady u cannot yield this result. With u constant the equilibrium at G_1^*, E_1^* is unstable, for the slope wN/uR will be changing.

$$\begin{aligned}(wN/uR)^* &= w^* + N^* - R^* \\ &= E_1^* + N^* - Q^* + G_1^* \\ &= G_1^* > 0\end{aligned} \tag{14}$$

so that E_1^* increases and G_1^* diminishes, until in steady-growth equilibrium,

$$G^* = 0, \quad E^* = E_2^* \tag{15}$$

That is, as before, all technical progress is labor augmenting. Not pollution R, but pollution per unit of output G is stabilized. However, output is growing (at a rate $N^* + E_2^*$), and pollution is growing *pari passu* with output. This is, to be sure, an improvement over G_0^*, E_0^*, but it is hardly satisfactory.

If R is to be stabilized at a given level, $\overline{R}$, then R will be essentially playing the part of land in a three-factor land, labor, capital model. Such a model, with an innovation possibility frontier, has been explored (see the author, 1970; Reichenbach). In such a model the price of the fixed factor must rise (see the author, 1970). Indeed, it is required that $u^* = E^* + N^* = Q^*$. It is indeed the given (but nonzero) price of environmental amenity which stabilizes growth at $G^* = 0$, just as the constant price of capital (due to proportionate saving) so stabilizes it in the two-factor model. If indeed, $u^* = E^* + N^* = Q^*$ and the initial value of u is large enough that $G_1^* = E^* + N^* = u^*$, then

$$(wN/uR)^* = w^* + N^* - u^* - R^* \tag{16}$$

$$R^* = Q^* - G_1^* = 0 \tag{17}$$

$$\begin{aligned}(wN/uR)^* &= w^* + N^* - u^* \\ &= w^* + N^* - w^* - N^* \\ &= 0\end{aligned} \tag{18}$$

Such a growth equilibrium is steady. Moreover, it is stable (see Reichenbach).

This has some importance for environmental policy. The most common proposal to deal with the destruction of common environmental amenities is that a tax should be placed either on pollution or on polluting activities. It seems unlikely, however, that pollution taxes would have the kind of automatic upward flexibility required to stabilize pollution in a technically dynamic economy. The proposal to establish a market for pollution quotas would have the needed upward flexibility, as would William Baumol's proposal. Despite the well-known shortcomings of quantitative regulation, such regulation might be a politically feasible "second best" solution, taking the implications of this model into account.

Baumol's proposal proceeds from the assumption that the optimal levels of pollution will be very difficult to discover in practice. Thus, the targets for each pollutant are determined politically and may not be optimal. Once the targets are determined, pollution taxes are adjusted by trial and error to the levels which induce restriction of pollutants to the target levels. Although the pollution levels attained may not be efficient, the reductions of pollution are efficiently distributed among potential

[5]While the research leading to this paper was underway, I reviewed a preliminary version of J. Kirker Stephens' paper, which covered some of the points above. My own back-of-the-envelope version closely paralleled Stephens' but was far more preliminary, and I have profited from his formulation. What follows, however, diverges from Stephens.

polluters, since all potential polluters face the same price of pollution. This avoids some of the obvious shortcomings of detailed quantitative regulations while securing the advantage of an automatic and appropriate rate of increase of the price of pollution.

We should observe that there can be no gain in G^* without a sacrifice in the growth rate of the standard of living, which is indicated by $E^* = Q^* - N^*$, the growth rate of income per unit of labor. The exact quantitative dimensions of the sacrifice required cannot be known.

Thus, the innovation-possibility frontier approach leads to the following conclusions. Where technology is flexible, a zero price of environmental amenity will indeed induce a bias in technical progress toward the destruction of the environment. This may well lead to an increase in pollution out of proportion to the increase in production. A stable positive price of pollution will eventually stabilize pollution per unit of output but will not stabilize the quantum of pollution in a growing economy. To stabilize pollution in a growing economy, it is necessary that the price of pollution increases at the same rate as output, if the innovation-possibility frontier model is descriptive of reality. However, this strong result does not hold true without modification in the more general model to which we now turn.

II

In a criticism of the innovation-possibility frontier approach, Nordhaus proposes an alternative and more general approach. Nordhaus defines an "isotech" as the set of all techniques which can be attained at the same given cost. The family of isotechs defines the technological opportunities available at a particular time. This approach has at least the following advantages over the innovation-possibility frontier approach as that approach exists in the literature: 1) technical progress is not supposed costless, 2) the overall rate of technical progress may be endogenously determined with reference to cost,[6] 3) it is independent of the steady-growth, factor augmenting technical progress assumptions which limit the generality of the innovation-possibility frontier approach.

Let us consider the problem of designing a new sort of plant for the purpose of supplying a particular product. The price of the product and the quantity demanded are, for simplicity, taken as given; units are normalized so that the price of the product is one. The characteristics of the new plant are expressed as

n = labor per unit of output
k = capital per unit of output
v = pollution per unit of output
Q = anticipated flow of sales

The plant[7] will be a one-hoss shay; its durability is given and denoted by m (it is assumed that the usefulness of the research is limited to the life of the plant on the hypothesis that, once worn out, the plant will be replaced by another of still later and better technological vintage). The value of the *plant* is the demand price,

$$(19)\quad \int_0^m e^{-rt}(Q - wnQ - uvQ)\,dt = Q\int_0^m e^{-rt}(1 - wn - uv)\,dt$$

where r, u, and w are as before. However, the value of the *design* will be the demand price minus the capital cost of the new plant:

$$(20)\quad \int_0^m e^{-rt}(Q - wnQ - uvQ)\,dt - kQ = Q\left[\int_0^m e^{-rt}(1 - wn - uv)\,dt - k\right]$$

[6]These difficulties are not inherent in the approach, however. My 1978 paper extended the innovation-possibility frontier approach to include development cost and the endogenous determination of the innovation-possibility frontier.

[7]It is not meant to imply that there would be only one *factory* built to a design, but a complex of factories of a number large enough to supply output Q.

The isotech will now be an array of (k, n, v) triplets which can be attained at the same cost; the zero isotech defines the neoclassical production function, which may or may not allow the smooth substitution of factors (compare Becker). In passing, we have implicitly imposed the restriction that this production function will display constant returns to scale, in the assumption that k, n, v are given regardless of Q. We may express the family of isotechs as an expenditure function,

$$E = g(k, n, v) \tag{21}$$

with $g(k_o, n_o, v_o) = 0$ where k_o, n_o, v_o are the contemporaneous values of k, n, v.

We proceed to explore the properties of a maximum of (20), subject to the constraint that

$$g(k, n, v) = E_o \tag{22}$$

Since the instrument variables k, n, and v are constants for the term of integration, the methods of the calculus of variations are not needed (however, compare the author, 1974b). The Lagrangian function is

$$L = Q\left[\int_0^m e^{-rt}(1 - wn - uv)dt - k\right] + \lambda(E_o - g(k, n, v)) \tag{23}$$

The necessary conditions for a maximum are

$$\frac{\partial L}{\partial n} = -Q\int_0^m e^{-rt}w(t)dt - \lambda\frac{\partial g}{\partial n} = 0 \tag{24a}$$

$$\frac{\partial L}{\partial v} = -Q\int_0^m e^{-rt}u(t)dt - \lambda\frac{\partial g}{\partial v} = 0 \tag{24b}$$

$$\frac{\partial L}{\partial k} = -Q - \lambda\frac{\partial g}{\partial k} = 0 \tag{24c}$$

We may denote

$$W = \int_0^m e^{-rt}w(t)dt \tag{25a}$$

as the wage per person over the entire life of the plant and

$$U = \int_0^m e^{-rt}u(t)dt \tag{25b}$$

as the total payment of pollution penalties over the life of the plant. In case w and u are constant over the life of the plant, we have

$$W = w\int_0^m e^{-rt}dt \tag{26a}$$

$$U = u\int_0^m e^{-rt}dt \tag{26b}$$

where

$$\int_0^m e^{-rt}dt$$

is, of course, the discounted present value of an annuity of one dollar for m years.

These conditions yield the following optimum rules in the familiar form:

$$U/W = (\partial g/\partial v)/(\partial g/\partial n) \tag{27a}$$

$$U = (\partial g/\partial v)/(\partial g/\partial k) \tag{27b}$$

$$W = (\partial g/\partial n)/(\partial g/\partial k) \tag{27c}$$

We can express these rules in diagrams of a familiar sort. For simplicity, however (and to avoid drawing three-dimensional diagrams), we suppress the variable k and consider only n, v, and equation (27a). Suppose, to begin, that $u = 0$ throughout the lifetime of the plant. Then $U/W = 0$. This is shown in Figure 2, where the "expansion" path aa' is the locus of all optima with $(\partial g/\partial v)/(\partial g/\partial n) = 0$. Any positive level of technical effort, such as isotech 1, will result in an increase in pollution per unit of output. A larger development effort, such as isotechs 2 and 3, will yield still more pollution per unit of output.

If, however U is positive and large enough, and W independent of the size of the development effort, then we have the contrasting possibility shown in Figure 3. The expansion path aa' shows decreasing pollution per unit of output, which may or may not be sufficient to offset growing output in a growing economy. If U is sufficiently large, it would seem that pollution

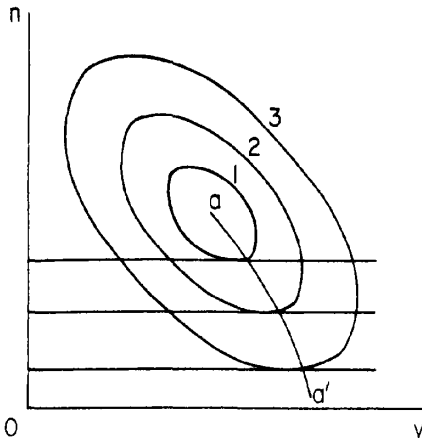



FIGURE 2

FIGURE 4

could be stabilized in a growing economy. This is, however, superficial. The assumption that W is independent of the size of the development effort is artificial. It would seem that a consistently larger development effort would, over time, result in a faster rise of wages because of the faster rise in labor productivity and so in labor demand One of the advantages of steady-growth models, which this approach gives up along with their simplifying assumptions, is that the steady-growth models take interdependencies of this kind into consideration.

To anticipate the effect of aggregate production on aggregate pollution in the terms of the model, we must take into account the impact of aggregate development effort on the aggregate demand for labor. No attempt will be made to do this formally. Let us consider the isotechs in Figure 4 to be aggregate ones, n to be the aggregate labor-output ratio, and v the aggregate pollution-output ratio. (It is not clear that such aggregate isotechs need exist, so this is just the kind of simplifying assumption we would very much like to avoid, if we could.) Each individual technical development manager may play the part of a perfect competitor, taking W as parametric and making decisions in accordance with (27a) above. (Here we have another fragment of the steady-growth general competitive equilibrium simplifying assumptions the isotech approach was partly meant to avoid,[8] and implicit in the mathematical model from the

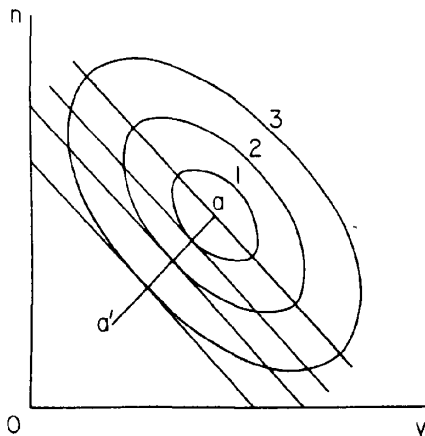
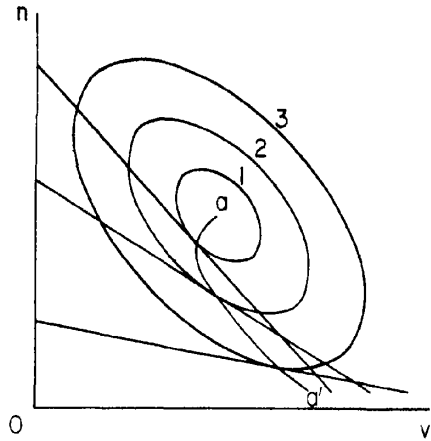



FIGURE 3

[8]Clearly, a labor-monopsony maximum would differ from equations (23) (27) above as would a product-market monopoly model However, it is not clear that a monopsony model would illuminate the issue considered in the text at this point, for we must presume that the elasticity of supply to a sector would be greater than that overall, even if it is not infinite to a sector Moreover, it seems likely that the final results would not be affected by monopsony in the different sectors, provided that technical development would not affect the *degree* of monopsony, that is, the elasticity of labor supply (or the degree of monopoly, similarly defined) The more interesting possibility, indeed, is that technical development *might* affect the degree of monopoly or monopsony This is, however, beyond the scope of this paper

beginning.) Then a larger aggregate development effort would result in a higher W, and with U given, this would lead to a pollution bias which, for a large research effort like isotech 3, approaches the same limit as we observed in Figure 2. To secure the antipollution bias shown in Figure 3, it is necessary again that U and hence u be flexible upward. With rising wages in a growing economy, u must be rising too and (this model suggests) at about the same rate as wages.

Thus, the two models suggest similar conclusions for environmental policy. First, zero pricing of environmental amenities will indeed produce a bias toward environmental pollution, so that pollution increases more rapidly than production. Second, if pollution is to be stabilized in a growing economy, the price of pollution must not only be positive but flexible upward and must rise at least *pari passu* with labor productivity and wages.

III

The discussion of the last section contains some tacit assumptions which should not go unmentioned. The equations are, of course, descriptive of interior maxima and describe only the necessary conditions for such maxima. The convexity of the isotechs and of the expenditure function, that is, the sufficient conditions for a minimum, are not self-evident and are assumed. In addition, however, the conclusions of the previous sections depend on the forms of Figures 1, 2, 3, and 4. Other forms are possible, and the purpose of this section is to point out these other logical possibilities in order to underscore the presence in the previous section of some powerful, if plausible, positive hypotheses about the nature of technological opportunities.

Consider first Figure 1. Had the inflection point at G_0^*, E_0^* been drawn in the first quadrant, that is, with G_0^* positive, the outcome would have been quite different. First, the bias with a zero price of environmental amenities would have been toward decreasing pollution per unit of output. Second,

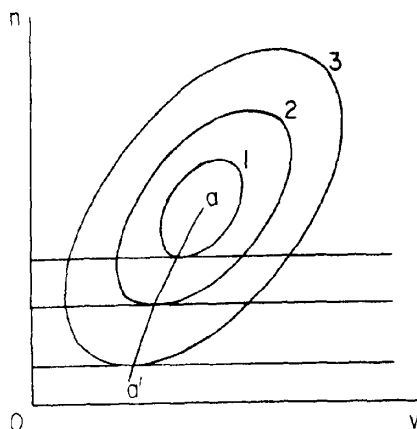



FIGURE 5

no steady-growth equilibrium would then have been possible with a positive pollution price. Clearly, in drawing Figure 1 as we did, we imposed a positive hypothesis of considerable strength and importance.

Second, consider Figures 2, 3, and 4. There the isotechs are drawn roughly elliptical[9] with the major axis a line of negative slope. Had the major axis been a line of positive slope, again, an altogether different practical conclusion would have been drawn. This is illustrated in Figure 5, where, despite the zero price of environmental amenities, technical development reduces pollution per unit of output. Here again, the figures are drawn in such a way as to introduce a powerful hypothesis.

Ultimately, the issue between Figures 2 and 5, for example, is an empirical one. However, the hypothesis tacit in Figures 1–4 is actually a very plausible one, if it is viewed as an extension of the neoclassical production function model.

Both the model of Section I and that of Section II are intended as generalizations of the neoclassical production function. The neoclassical model assumes that costless substitutions among factors may be made

[9]No free disposal is assumed, by contrast with Nordhaus. With respect to pollution, free disposal would be troublesome. It sometimes happens that a change of circumstances will render saleable a joint product which was formerly a polluting waste.

with positive marginal productivities over a wide range of factor ratios. The model of Section II, by contrast, assumes that all factor substitutions have a positive cost of transition. This renders the marginal productivity of a factor undefined. However, the marginal rate of substitution (at a given cost of transition, i.e., given E) is defined, and it is precisely $-(\partial E/\partial n)/(\partial E/\partial v)$. The hypothesis that both marginal productivities are positive implies that the marginal rate of substitution is negative. We extend this hypothesis by assuming that $-(\partial E/\partial n)/(\partial E/\partial v)$ is negative for small departures from n_0, v_0. This implies that isotechs are downward sloping for (n, v_0), and (n_0, v) as in Figures 2–4.

Furthermore, this reasoning applies to Figure 1 as well, since the innovation-possibility frontier can be considered as a transformed isotech of special form. At any moment of time, for example, $E^* = \dot{E}/E$, $F^* = \dot{F}/F$, and $G^* = \dot{G}/G$ define a set of E, F, G, attainable within a specified period of time at zero cost or indeed at any higher cost. Moreover, $E = 1/n$, $F = 1/k$, and $G = 1/v$. Thus inverted, the innovation-possibility frontier becomes a degenerate family of isotech curves (which is not, however, stable over time), and the argument above can be applied *pari passu*

The figures and the mathematical models contain another important tacit assumption, although on it the models differ. The difference in turn explains the difference in the results, in that the model of Section II relates optimum to relative factor *prices*, whereas that of Section I relates optimum to relative factor *expenditures*. The difference is that in Section II the isotechs are tacitly supposed to be stable functions of *absolute* deviations from base productivities, while in Section I the innovation possibility is defined by a stable function of *relative* deviations from base productivities.

Suppose that, in place of (21), we write

$$E = g'\left\{\frac{k}{k_o}, \frac{n}{n_o}, \frac{v}{v_o}\right\} \tag{28}$$

with $g'(1, 1, 1) = 0$. Then

$$\begin{aligned} \partial E/\partial k &= (1/k_o)(\partial g'/\partial k) \\ \partial E/\partial n &= (1/n_o)(\partial g'/\partial n) \\ \partial E/\partial v &= (1/v_o)(\partial g'/\partial v) \end{aligned} \tag{29}$$

and (27a) becomes

$$U/W = (n_o/v_o)(\partial g'/\partial v)/(\partial g'/\partial n) \tag{30}$$

or

$$v_o U/n_o W = (\partial g'/\partial v)/(\partial g'/\partial n) \tag{31}$$

and the slopes of the "budget lines" in Figures 2, 3, and 4 become the anticipated relative *expenditures* if factor proportions remain unchanged. The isotech approach then yields substantially the same result as does the innovation-possibility frontier approach (compare Becker).

Can any case be made for either (21) or (28) against the other? In other words, can we defend the notion that technical transition cost is an invariant function of absolute deviation from current productivities and not the relative deviation or vice versa? It is generally agreed that E depends, not only on the deviations $k - k_o$, $n - n_o$, $v - v_o$, but also on the previous history of productivity advance (see Nordhaus). The explanation underlying (28) is a very simple hypothesis as to how this might occur. Probably things are not so simple, but it merits consideration. Undoubtedly other, equally simple hypotheses can be incorporated in the form of $g(\ \ )$ in (21). In any case, the issue is an empirical one and at this stage a very murky one.

The primary point, however, is that the results of Sections I and II are not, appearances to the contrary, in contradiction. Rather, the different results merely reflect different mathematical forms, for if $g(k, n, v) = g'(k/k_o, n/n_o, v/v_o)$, the two optima (27) and (31) correspond precisely (compare Syed Ahmad).

Both models, then, seem to be reasonable extensions of the neoclassical model, in the light of the tacit assumptions which underlie the diagrams and mathematical models. That the model of Section II is more general has been noted. This generally leads to

a different mathematical form, but the results, though different in appearance, encompass those of Section I as a special case

IV

I have explored two models of the impact of economic variables on trends in the productivity of labor and capital and pollution per unit of output. On the assumptions that technology is flexible in response to economic forces and that reductions in pollution per unit of output must be traded off against improvements of productivity, the intuitive notion is confirmed that technical progress will be biased toward use of unpriced environmental amenities. In that case, pollution will increase more than in proportion with output. However, pollution cannot be stabilized in a growing economy by a positive *and constant* pollution price. Rather, the pollution price must rise at least proportionately with the productivity of labor, that is, the wage. It must rise even more rapidly in some cases. Depending on the characteristics of the model, it may be necessary for the pollution price to rise proportionately with the total wage bill. In any case, it is necessary that the pollution price be flexible upward, if pollution is to be stabilized. One may doubt that direct legislation of such penalties as taxes will provide for appropriate flexibility. The proposal for markets in pollution rights and the Baumol proposal do provide for appropriate flexibility. However, quantitative regulations may be the best politically feasible approach to environmental management, when these long-run tendencies are taken into account.

REFERENCES

S. Ahmad, "On The Theory of Induced Invention," *Econ. J.*, June 1966, *76*, 344–57.

W. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–22.

Gary Becker, *Economic Theory*, New York 1971.

H. Daly, "The Steady-State Economy: Toward a Political Economy of Biophysical Equilibrium and Moral Growth," in his *Toward a Steady-State Economy*, San Francisco 1973.

W. Fellner, "Two Propositions in the Theory of Induced Innovations," *Econ. J.*, June 1961, *71*, 305–08.

———, "Measures of Technical Progress in the Light of Recent Growth Theories," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1073–99.

John R. Hicks, *The Theory of Wages*, 2d ed., London 1963.

C. Kennedy, "Induced Bias in Innovation and the Theory of Distribution," *Econ J.*, Sept. 1964, *74*, 541–47.

R. A. McCain, "Land in Fellner's Neoclassical Model of Economic Growth: Comment," *Amer Econ. Rev.*, June 1970, *60*, 495–99.

———, (1974a) "Further Comment of 'Smoothing' by Output Variation," *Quart. J. Econ.*, Aug. 1974, *88*, 496–98.

———, (1974b) "Induced Bias in Technical Innovation Including Product Innovation in a Model of Economic Growth," *Econ J.*, Dec. 1974, *84*, 959–66.

———, "Economic Growth with an Endogenous Innovation-Possibility Frontier," unpublished paper, 1978.

W. D. Nordhaus, "Some Skeptical Thoughts on the Theory of Induced Innovation," *Quart. J. Econ.*, May 1973, *88*, 208–19.

H. Reichenbach, "Boden in Einer Wachsenden Wirtschaft," unpublished doctoral dissertation, Institut fur Agarpolitik Und Marktlehre, Christian-Albrechts Universität Kiel 1974.

Jacob Schmookler, *Invention and Economic Growth*, Cambridge 1966.

J. K. Stephens, "A Relatively Optimistic Analysis of Growth and Pollution in a Neoclassical Framework," *J. Environ Econ Manage.*, Aug. 1976, *3*, 85–96

R. Solow, "A Contribution to the Theory of Economic Growth," *Quart J Econ.*, Feb. 1956, *70*, 65–94.

Empirical Tests of the Life Cycle Hypothesis

By BETSY BUTTRILL WHITE*

Since 1954, when Franco Modigliani and Richard Brumberg published the seminal article on the subject, the life cycle hypothesis has become a well-established model of personal consumption and saving. The model has been subjected to numerous empirical tests whose results allege to be supportive of the hypothesis.[1] This paper outlines a general critique of previous regression analyses of the model and presents a new test of the model. The test demonstrates that the model is lacking as an explanation of aggregate personal saving in the United States.

The methodology employed in this paper, simulation analysis, takes a set of assumptions and calculates the quantitative implications of these assumptions. That is, it answers the question, what are the values of the variables in question when the economy and the economic actors behave exactly according to the assumptions employed? The major advantage of simulation analysis over regression analysis in testing the life cycle model is that it obviates the aggregation and data problems which generally vitiate regression results.

In testing the model, several sets of simulations are performed using various assumptions about the lifetime income and consumption streams of individuals. These simulations calculate the aggregate personal saving implied by the life cycle model under the assumptions employed. The results are strengthened by the use of actual aggregate income and demographic data. Although the simulations are based upon the specific micro foundation assumed by the authors of the life cycle model, the results of the simulations are sufficiently strong to apply to all theories of consumption which propose that people abstain from current consumption in order to provide solely for future consumption and that this sort of abstention explains the totality of observed personal saving.

The major conclusion drawn from this analysis is that saving for future consumption does not account for the totality of observed aggregate personal saving. For a wide range of parametric values, the simulated values of aggregate saving fall significantly short of the observed levels. At best, the simulated values are about 60 percent of the observed values.

I. Limitations of Regression Analyses of the Life Cycle Model

The ultimate goal of empirical testing in economics is to determine how well a theoretical model corresponds to the reality of an economic situation. In order to be fruitful, an empirical test must differentiate the empirical implications of one theory from another and the significance of the results of the tests must be determinant. In regression analysis, the regression equation should represent only one theoretical model. If the estimating equation can be derived from either of two models, regression analysis using a given set of data cannot tell us which, if either, of the models is supported by the data

In this respect, there are two major interrelated problems in using regression analysis to test the life cycle model. They deal with the aggregation procedure used in deriving the aggregate function to be tested and with

*Economist, Federal Reserve Bank of New York This article is based on my Ph D dissertation submitted to Stanford University I would like to thank Duncan K Foley for his guidance and insightful comments The views expressed in this article are mine and do not necessarily reflect the views of the Federal Reserve Bank of New York

[1] See, for example, Albert Ando and Modigliani (1957, 1963), Brumberg, M J Farrell, Malcolm Fisher, Michael Landsberger, Keizo Nagatani, and James Tobin

the data used in the regression. Given adequate data on the age, lifetime income, and net worth of individuals, the life cycle model could be tested at the micro level. However, since such data are not available at this time, aggregate data have been used to test the model. Unfortunately, in order to aggregate individual consumption functions into an aggregate consumption function which is estimable, several rather unwarranted assumptions are made. The usual aggregation procedure produces a function of the general form:

$$C_t = b_1 Y_t + b_2 A_t \tag{1}$$

where C_t = aggregate consumption at time t
Y_t = aggregate income at time t
A_t = aggregate net worth at time t
b_1, b_2 = coefficients to be estimated

This equation is not representative of the life cycle model and, even if it were, it is also representative of other consumption models. For example, this equation is representative of the Keynesian model when capital gains are considered important

One of the basic assumptions[2] of the life cycle model implies that the individual's utility function takes the form.

$$U = \sum_{t=1}^{T} u(c_t)(1+p)^{1-t} \tag{2}$$

where $u(c_t) = (1/(1-d))c_t^{1-d}$ + a constant ($d > 0$ and $d \neq 1$)
$u(c_t) = \ln c_t$ + a constant ($d = 1$)
T = years in the individual's economic life
c_t = flow of consumption services in time period t
p = rate of pure time preference
$-d$ = elasticity of marginal utility with respect to consumption[3]

Three other assumptions[4] of the model specify the individual's lifetime budget constraint as

$$\sum_{i=1}^{T_j} c_i^j(1+r)^{1-i} = a_t^j + \sum_{i=1}^{T_j} y_i^j(1+r)^{1-i} \tag{3}$$

where T_j = years remaining in the jth individual's economic life in time period t
c_i^j = consumption by individual j in year i of his remaining life
y_i^j = earned income of individual j in year i
r = real rate of return at which the individual can borrow or lend without limit
a_t^j = net worth of individual j in time period t (zero at the beginning of economic life)

Maximizing equation (2) subject to the constraint (3), renders the rate of growth of individual consumption along the optimal consumption path as

$$\frac{c_{t+1} - c_t}{c_t} = \left(\frac{1+r}{1+p}\right)^{1/d} - 1 \tag{4}$$

Using expression (4) in equation (3), and solving for c_t^j, the individual's consumption function becomes

$$c_t^j = \left[\sum_{i=1}^{T_j} y_i^j(1+r)^{(1-i)} + a_t^j\right] \Big/ \Delta_j \tag{5}$$

$$\text{where} \quad \Delta_j = \sum_{i=1}^{T_j}\left(\frac{1+r}{1+p}\right)^{(i-1)/d}(1+r)^{(1-i)}$$

The aggregate consumption function de-

[2]Modigliani and Brumberg assume that the individual's utility function is "such that the *proportion* of his total resources that an individual plans to devote to consumption in any given year of his remaining life is determined only by his tastes and not by the size of his resources" (1969, p. 106). Both Menahem Yaari and Earl Thompson provide proofs that such an assumption implies the utility function of equation (2)

[3]Alternately, $(-1/d)$ is the elasticity of substitution between consumption in one period and the next.

[4]Ando and Modigliani (1963) assume 1) "The individual neither expects to receive nor desires to leave any inheritance" (p. 56), 2) "The rate of return on assets is constant and is expected to remain constant" (p. 59), and 3) Capital markets are perfect.

rived from the life cycle model is merely the sum of all the individual consumption functions. This aggregate function is given in equation (6).

$$(6)\quad C_t = \sum_{j=1}^{P} c_1^j = \sum_{j=1}^{P} \frac{\sum_{i=1}^{T_j} y_i^j (1+r)^{(1-i)}}{\Delta_j} + \sum_{j=1}^{P} \frac{a_1^j}{\Delta_j} = \frac{Y_t}{\sum_{j=1}^{P} \Delta_j} + \sum_{j=1}^{P} \left[\frac{\sum_{i=2}^{T_j} y_i^j (1+r)^{1-i}}{\Delta_j} \right] + \sum_{j=1}^{P} \frac{a_1^j}{\Delta_j}$$

where C_t = aggregate consumption in time period t
Y_t = aggregate income in time period t
P = population size in time period t

When equation (1) is used in a time-series analysis instead of the true equation, equation (6), the effects of changes in the age distribution, the age distribution of income, and the age distribution of net worth are completely ignored, and it is these factors which empirically differentiate the life cycle model from other models of consumption and saving.[5] Thus, one must look beyond the technique of regression analysis in testing the life cycle model.

II. A Simulation Analysis of the Life Cycle Model

Instead of estimating the coefficients of the aggregate saving equations, the simulations reported below calculate the levels of aggregate saving suggested by the assumptions of the model and the sensitivity of these levels to various changes in these assumptions. The test of the model is the comparison of these simulated values of saving to the actual values of saving during three particular years (1953, 1959, and 1964).

A. *The Mechanics of the Simulations*

The procedure for calculating the simulated value of aggregate saving is conceptually rather simple.[6] Using equations (3) and (4), the individual's optimal consumption path is

$$(7)\quad c_t = c_1 \left(\frac{1+r}{1+p} \right)^{(t-1)/d}$$

$$\text{where}\quad c_1 = \frac{\sum_{i=1}^{T} y_i (1+r)^{1-i}}{\sum_{i=1}^{T} \left(\frac{1+r}{1+p} \right)^{(i-1)/d} (1+r)^{(1-i)}}$$

and where the variables are defined as in Section I. The level of this optimal consumption path depends on the present value of the hypothesized lifetime income stream. The authors of the life cycle model assume that all individuals have the same income and that they all expect to receive equal incomes t years from now (provided they all are still working at that time). In the following simulations, several different lifetime income streams are hypothesized to test the sensitivity of the model to this assumption.

Given a lifetime income stream, the individual's net worth and current saving can be derived using equation (7). The individual's net worth a_t is given by

[5]Given sufficiently disaggregated cross-section data, one could use an equation similar to equation (1) to estimate b_1 and b_2 for groups with the same age, expected income and net worth The life cycle model could be tested by comparing these estimated coefficients to values suggested by the model An attempt at such an analysis was made by Harold Watts See the author (ch. 2) for a discussion and critique of this and other attempts to use cross-section data in testing the model

[6]The simulations were performed using a Fortran computer program written by me and based on equations (7) (10) The program calculates for an individual in each age group in the economy a lifetime income stream; a lifetime optimal consumption path, a lifetime saving path; and net worth positions for each time period. In calculating aggregate saving, the program multiplies the simulated saving of an individual in a given age group by the actual number of people in that age group and then sums this group saving over all age groups. A program listing and description are available from the author upon request.

$$a_t = \sum_{i=1}^{t-1} (y_i - c_i)(1 + r)^{t-1-i} \tag{8}$$

His current saving is simply the difference between his total current income and his current consumption:

$$s_t = y_t + a_t r - c_t \tag{9}$$

where $a_t r$ is the individual's interest income. Finally, aggregate personal saving is calculated by summing across all individuals in the economy:

$$S_t = \sum_{j=1}^{P} s_t^j \tag{10}$$

where S_t is current aggregate personal saving and P is the number of consumer units in the economy.

The earned income concept suggested by the life cycle model is disposable labor income plus contributions to social security and pension funds. However, due to the lack of such data, total disposable personal income is used here. To the extent that this variable overestimates disposable labor income by including nonlabor income, the following simulations overestimate the level of saving suggested by the life cycle model.

The simulations reported here use age distributions and aggregate income for the years 1953, 1959, and 1964. The year 1953 was chosen initially because it is the post-World War II year in which the unemployment rate was the lowest. However, as a result of Korean War expenditure, the economy in that year was overstimulated. The cyclical expansion peaked in that year. According to the life cycle model, during periods of unexpectedly high income, saving also should be relatively high in order to carry the consumption benefits of the unexpected income over to future periods. The year 1959 was chosen as an approximation to a situation in which cyclical fluctuations play little or no role. The first quarter of 1958 marked the trough of a business cycle. The following business upturn, of which 1959 was a part, was relatively short and undramatic in its effects. Thus, the economy was neither at a trough nor at a peak and the expansion that occurred was not as dramatic as that in early 1953. The year 1964 was not dissimilar to the year 1959 in that it was the middle of an expansion phase, but the phase was much longer than the one following the 1958 trough.

For the following simulations, it is assumed that individuals enter the labor force at age twenty, retire on their sixty-fifth birthday, and die on their seventy-fifth birthday. It is also assumed that aggregate income is earned by individuals between twenty and sixty-five years of age, and that each individual of a given age earns the same income. The assumptions of a 45-year earning span and a 10-year retirement period are used because they more closely approximate reality than Ando and Modigliani's (1963) assumption of a 40-year working span and 10-year retirement period.[7]

The lifetime income stream for each age group in the economy is calculated such that the sum of current income of all individuals in the economy is equal to actual observed disposable personal income in the year in question.[8] Given these income streams, the lifetime consumption stream for each age group is derived using equation (7). Because these consumption paths are dependent upon the relationship between the rate of return r and the individual's rate of time preference p, and the elasticity of marginal utility $-d$, a wide range of values for these parameters is used in each set of simulations. Given lifetime income and consumption streams, the lifetime saving and net worth paths are calculated for each age group. The saving paths are constrained such that the individual's net worth at time of death is zero. (This results from

[7]Although the life expectancy of twenty-year olds in 1959–61 was only 71.8, the life expectancy of forty-year olds in that period was 74.4. Further, the use of sixty-five as the retirement age is more realistic than sixty since the labor force participation rate in 1960 was 85.2 percent for males between fifty-five and sixty-five years of age and 32.2 percent for males sixty-five and over. (See U.S. Bureau of Census, *Statistical Abstract*, 1973, p. 57, and 1975, p. 44.)

[8]Disposable personal income figures are from the *Survey of Current Business*.

TABLE 1—SIMULATED AGGREGATE SAVING
(Income and Consumption Constant)

Year	Actual Saving[a]	Interest Rate				
		−.03	00	.03	.05	.07
1953	61.4	17.6	15.5	11.9	9 5	7 5
1959	72.3	30.4	28.9	25.0	21.9	18 9
1964	91 7	34 1	31 8	27.2	23 7	20 5

[a]All figures are in billions of 1958 dollars The saving figure for 1953 is from Raymond W. Goldsmith, et al , p 361 The saving figures for 1959 and 1964 are from Board of Governors of the Federal Reserve System, p 2 The saving figure includes gross purchases of consumer durable goods, and thus is not the same personal saving concept reported by the Department of Commerce The details are outlined in the cited sources

the model's assumption of no bequests.) At any point in time t during the individual's economic life, the individual's net worth and current saving are given by equations (8) and (9), respectively In order to calculate aggregate personal saving (using equation (10)), the actual age distribution and population size for the year in question are used Thus, the mechanics of the simulations are derived directly and exclusively (except for the lengthened earning span) from the original assumptions of the model.

B. *Aggregate Saving Explained by the Original Model*

The first set of simulations presented here employs the original assumptions of the model with only slight modification. That is, these simulations assume that desired consumption paths are horizontal,[9] everyone in the earning span receives the same earned income, and earned income is constant over the individual's earning span. The results are given in Table 1 The negative rate of interest is used for the sake of completeness. It has been suggested that in inflationary periods, the average saver, holding much of his portfolio in fixed nominal valued assets, might experience a negative real return on his assets. It is clear that the life cycle model in its purest form does not explain the totality of aggregate household saving. For every rate of interest, the results fall substantially short of the actual levels observed. This is true for each of the three years independent of the business phase represented by the year.

C. *Sensitivity of the Model to Changes in Assumptions*

The authors of the life cycle model claim that their assumptions regarding lifetime income and consumption paths are not essential to the model. The remaining simulations in this paper test a wide variety of desired consumption paths and income streams to determine the degree to which the results of this original set of simulations depend upon the model's original assumptions. It will be seen that changes in the assumptions generally lead to lower, rather than higher, simulated saving figures.

1. *Increasing Income*

The next set of simulations calculates the level of saving suggested by the model when the individual's income stream is upward sloping. Specifically, income is assumed to increase by the same dollar amount each year from age twenty to fifty-four and then to remain constant until retirement. This shape for the lifetime income stream is sup-

[9]Ando and Modigliani assume that the consumer "plans to consume his total resources evenly over the remainder of his life span" (1963, p 59) However, using equation (7), this is equivalent to the assumption that the rate of pure time preference is equal to the real rate of interest

TABLE 2 — SIMULATED AGGREGATE SAVING
(Linearly Increasing Income and Constant Consumption Paths)

		Interest Rate				
Year	Actual Saving[a]	−.03	.00	.03	.05	.07
1953	61.4	4.9	−3.6	−11.0	−14.3	−16.1
1959	72.3	24.3	14.1	3.5	−2.2	−6.2
1964	91.7	24.9	14.4	3.6	−2.0	−6.1

[a]See Table 1

ported by data from surveys of family income. (For example, see Harold Lydall, p. 143, and Dorothy Projector, p. 15.) Assuming that twenty-year olds received the average of the income earned by those between fourteen and twenty-four in the years in question,[10] and then constraining the equation to assure that the aggregate income calculated in this manner will be equal to the actual aggregate disposable income in each of the years in question, income can be calculated for the individual at each age in his earning span using

$$y_i = y_0 + h(\text{age}_i - 20)$$

where h = a constant
age_i = age in year i for $i \le 54$ and 54 for $54 < i \le 64$

In these simulations, the consumption path again is assumed to be horizontal The results are presented in Table 2. The values of aggregate household saving were reduced (some even became negative!) by the introduction of an upward-sloping lifetime income stream. These lower values of saving result from the net debtor position of the younger people in the simulated economy. In anticipation of higher future incomes, these people consume more than they earn when they are young in order to maintain a constant consumption path over their lifetimes.

[10]Income figures for the fourteen to twenty-four-year age bracket are from U.S. Bureau of the Census, Technical Report No. 17, *Trends in the Income of Families in the U.S., 1947–1964*.

2. *Families*

It appears that Malcolm Fisher was the first to call attention to the absence of families in the life cycle model: "As presented, the theory might be described as the bachelor theory of saving" (p. 224). In reality, the average household size does not remain constant over its lifetime. In general, the family size increases as the head of the household grows older until about middle age, usually due to the growing number of children in the household. By the time the head of the household reaches middle age, children start to leave home and the family size diminishes. Therefore, it is not realistic to impute a desire for a constant level of consumption to households. In order to incorporate changes in family size in calculating the optimal lifetime consumption, the lifetime budget constraint must be changed to incorporate family size.

$$(3') \quad \sum_{i=1}^{T} c_i^e f_i (1+r)^{1-i} = \sum_{i=1}^{T} y_i (1+r)^{1-i}$$

where c_i^e = consumption per person equivalent in the ith year of the economic life of the head of the household
f_i = number of people in the household during the ith year of the economic life of the head of the household
T = number of years in the economic life of the head of the household

Using consumption per person equivalent

TABLE 3—SIMULATED AGGREGATE SAVING
(Linearly Increasing Income and Consumption Related to Family Size)

Year	Actual Saving[a]	Interest Rate −.03	.00	.03	.05	.07
1953	61.4	−6.1	−13.6	−18.4	−19.7	−20.0
1959	72.3	1.7	−8.2	−14.6	−16.8	−17.7
1964	91.7	3	−9.7	−15.8	−17.7	−18.4

[a]See Table 1.

in the utility function of the head of household, in a manner analogous to the procedure used above,

$$(7') \qquad c_t = c_1^o f_t = c_1^o \left(\frac{1+r}{1+p}\right)^{(t-1)/d} f_t,$$

where
$$c_1^o = \frac{\sum_{t=1}^{T} y_t (1+r)^{1-t}}{\sum_{t=1}^{T} f_t \left(\frac{1+r}{1+p}\right)^{(t-1)/d} (1+r)^{1-t}}$$

Since the data on f_t are given for families, the number of unrelated individuals in each group is subtracted out of the number of individuals for each age group. Consumption and saving for single individuals are calculated in the same manner as before. The remaining people in each age group are assumed to be married to people of their own age and to have children or other dependents. The profile of the number of people in each family over the lifetime is the same for all families. Ideally, one would like to assign to each cohort of families the number of children expected by that cohort to reflect the fact that the size of the average family has changed over time. However, since these data are not available, the 1955, 1959, and 1964 cross-section profiles of families are used to represent the individual family's lifetime profile.[11]

Admittedly, such a procedure might lead to erroneous results. However, it might be justified in this situation since such a profile should render a better estimate of the situation than the previously assumed constant level of consumption. Similarly, it is probably unrealistic to assume that unrelated individuals have always been and always plan to be unrelated individuals or to assume that they plan their consumption in the same way as a person who plans to be single over his entire economic life. However, this is exactly what the original life cycle model assumed for all individuals. Thus, although imperfect, perhaps this analysis brings us a step closer to reality.

The results of the simulations are given in Table 3. The lifetime income stream is the same as the one used in the previous set of simulations. The impact of family size is seen to be a decrease in aggregate household saving. For a given level of resources, lifetime consumption per capita must diminish when children are added. More is consumed during the years before retirement due to the larger family size, and therefore less can be saved for retirement years. Thus, the addition of the realism of families does not render the model better able to explain observed saving; on the contrary, the amount of saving explained decreases.

3. *Exponentially Growing Income*

The use of cross-section observations of income related to age to approximate an individual's lifetime income ignores the general growth in productivity which increases an individual's income independently of his age. Therefore, the next set of simulations assumes that the individual's income grows exponentially, rather than linearly, over his

[11]These data are taken from U.S. Bureau of the Census, *Current Population Report*, Series P-20, 1956, 1960, 1965.

TABLE 4 SIMULATED AGGREGATE SAVING
(Income Increasing Exponentially and Consumption Related to Family Size)

Year	Actual Saving[a]	Interest Rate -.03	.00	.03	.05	.07
Growth Rate of Income = .03						
1953	61.4	-7.0	-12.7	-15.8	-16.4	-16.1
1959	72.3	-1.4	-9.7	-14.6	-15.9	-16.1
1964	91.7	-1.3	-9.8	-14.5	-15.6	-15.8
Growth Rate of Income = .04						
1953	61.4	-13.0	-19.4	-22.4	-22.7	-22.0
1959	72.3	-6.8	-16.2	-21.3	-22.4	-23.3
1964	91.7	-6.8	-16.3	-21.0	-21.9	-21.7

[a]See Table 1

earning span. The inclusion of families remains. Again, the income received by individuals of age twenty is constrained to make aggregate income in the simulations equal to the observed aggregate disposable personal income in each of the three years. The results of these simulations are given in Table 4. This change in the shape of the individual's income stream does not affect the general result of negative saving resulting from an upward sloping income stream and from a constant per person consumption stream. Heavy borrowing in the household's early years keeps it in a net debtor position until the head of household is in his late fifties or early sixties.

4. *Increasing Per Capita Consumption*

The assumption that desired consumption per person equivalent is constant over the individual's or household's lifetime implies that the individual's rate of pure time preference is equal to the rate of interest. From equations (7) and (7'), it is clear that the sign of the rate of growth in per capita consumption along the optimal consumption path is determined solely by the relationship between the rate of time preference and the rate of interest. If the rate of time preference is less than (greater than) the rate of interest, optimal per capita consumption will increase (decrease) over time. Now let us assume that the individual wishes to have an upward-sloping consumption path. In Keynes' words, "it gratifies a common instinct to look forward to a gradually improving standard of life rather than the contrary, even though the capacity for enjoyment may be diminishing" (p. 108).

Given that individuals might desire an upward-sloping consumption path, what are reasonable rates of growth for desired consumption? It has been demonstrated elsewhere (see the author, pp. 63-65) that, for reasonable values of the consumption elasticity of marginal utility ($-d$ in equations (7) and (7')), rates of growth in consumption which are greater than the rate of interest generally imply negative rates of time preference. Negative rates of time preference imply that individuals are willing to forego a unit of current utility in return for less than a unit of future utility, that the present satisfaction derived from future consumption increases the further that consumption is into the future. Therefore, it seems most reasonable that cases of negative rates of time preference be omitted from consideration.

Table 5 presents the simulated values of aggregate personal saving for exponentially growing per person equivalent consumption for reasonable parametric values. Income grows exponentially as in the previous simulations. The implied rates of pure time preference are given for the special case of a logarithmic utility function ($d = 1$). For rates of growth of consumption less than .07, the values of saving are considerably below the actual values in each of the three

TABLE 5—SIMULATED AGGREGATE SAVING
(Income and Consumption Increasing Exponentially; $d = 1$)

	Interest Rate				
	-.03	.00	.03	.05	.07
I. Growth Rate of Income = .03					
a. Consumption Growth = -.03					
Implied time preference	.000	.031	.062	.082	.103
1953	-31.9	-34.2	-32.4	-30.1	-27.5
1959	-27.4	-31.7	-31.4	-29.9	-27.9
1964	-30.8	-33.2	-31.6	-29.5	-27.3
b. Consumption Growth = .00					
Implied time preference		.000	.030	.050	.070
1953		-12.7	-15.8	-16.4	-16.1
1959		-9.7	-14.6	-15.9	-16.1
1964		-9.8	-14.5	-15.6	-15.8
c. Consumption Growth = .03					
Implied time preference			.000	.019	.039
1953			9.5	7.5	5.8
1959			16.8	13.5	10.9
1964			16.2	12.4	9.5
d. Consumption Growth = .05					
Implied time preference				.000	.019
1953				29.0	28.0
1959				44.2	42.2
1964				41.8	38.9
e. Consumption Growth = .07					
Implied time preference					.000
1953					56.4
1959					86.7
1964					80.9
II. Growth Rate of Income = .04					
a. Consumption Growth = -.03					
Implied time preference	.000	.031	.062	.082	.103
1953	-39.5	-41.4	-38.7	-35.7	-32.5
1959	-34.2	-38.5	-37.6	-35.6	-33.0
1964	-38.0	-40.0	-37.6	-35.0	-32.2
b. Consumption Growth = .00					
Implied time preference		.000	.030	.050	.070
1953		-19.4	-22.4	-22.7	-22.0
1959		-16.2	-21.3	-22.4	-22.3
1964		-16.3	-21.0	-21.9	-21.7
c. Consumption Growth = .03					
Implied time preference			.000	.019	.039
1953			2.2	-.1	-1.8
1959			9.2	5.3	2.5
1964			8.7	4.4	1.4
d. Consumption Growth = .05					
Implied time preference				.000	.019
1953				20.3	18.7
1959				34.3	31.2
1964				32.1	28.3
e. Consumption Growth = .07					
Implied time preference					.000
1953					44.8
1959					71.9
1964					66.6

Note: Savings figures in billions of 1958 dollars.

years. In fact, for rates of growth below .03, aggregate saving is negative. Once again, when the growth rate in consumption is less than the growth rate in income, borrowing by the young in anticipation of higher future incomes leads to an overall negative level of saving.

The simulation using .07 for the growth rate in consumption renders good estimates of the 1953 saving when the growth rate in income is .03 and of the 1959 saving when the growth rate in income is .04. However, the estimates for all but one of the other years are too small. The exception is the year 1959, when income is growing at a rate of .03. In this case, the simulated value is 20 percent higher than the actual value. Thus, the simulations using this set of parameters are not consistent over the three years in their ability to account for aggregate saving. Furthermore, the use of a .07 growth rate in consumption is presented here for the sake of completeness, not for the sake of realism. A .07 rate of growth in desired consumption implies the real consumption services enjoyed at the end of life are 38.6 times as great as the real consumption services enjoyed at the beginning of the economic life, and this is clearly not representative of households in the United States during the period covered. In conclusion, then, the use of exponentially growing income and consumption does not render the simulation more effective in accounting for observed aggregate household saving.

5. *Age-Related Income Streams*

Thus far, it has been assumed that income and consumption streams of all individuals are identical. The last step toward realism to be taken in this paper makes the individual's income stream, and thus his consumption stream, dependent upon the calendar year in which he begins his economic life (the year in which he has his twentieth birthday). An individual's lifetime income stream will now explicitly reflect both secular (exponential) growth in income and growth in income resulting from the individual's accumulation of experience and expertise in his job. The cross-section profile of median incomes for different age groups in the three years under consideration were used to reflect the growth in income due to increases in experience and expertise.[12] For each individual the lifetime income stream is calculated by applying an exponential growth factor to this income profile to reflect the calendar year during which the individual reaches each age. This procedure makes the simulations more realistic in that it takes into consideration the fact that the lifetime real income of current old people is less than the lifetime real income of current young people and that therefore the lifetime saving will be different for the two age groups. Once again, the income profiles are constrained so that aggregate income in the simulations is equal to the actual level of income. The consumption paths of the individual are determined in the same way as in the previous set of simulations: family size is taken into consideration and optimal per person equivalent consumption grows exponentially.

The results of this last set of simulations are given in Table 6. The unrealistic rate of growth in consumption of .07 and the negative rate of interest are omitted as parameters as they do not add useful information. A lower exponential rate of growth in income is considered because the increase in income resulting from age now is calculated directly in the simulations. That is, the exponential rate of growth no longer subsumes both types of growth in the individual's income as it did in the simulations represented by Tables 4 and 5.

As might be expected, this set of simulations renders significantly higher levels of aggregate saving (when saving is positive) than the previous set of simulations. This is to be expected because now the retired people are dissaving less on average each year than the savers are saving. That is, the younger population looks forward to a higher level of lifetime consumption than

[12]Median income figures are from U.S. Bureau of the Census, *Current Population Report*, Series P-60, 1955; 1961, 1965

TABLE 6—SIMULATED AGGREGATE PERSONAL SAVING
(Income Stream Affected by Age and Secular Income Growth; Consumption Increasing Exponentially; $d = 1$)

	Interest Rate			
	.00	.03	.05	.07
I. Exponential Growth Factor = 1.02				
a Consumption Growth = .00				
Implied time preference	.000	.030	.050	.070
1953	−7.2	−16.1	−19.5	−21.1
1959	−14.0	−23.1	−25.9	−26.9
1964	−23.3	−32.5	−34.9	−35.2
b Consumption Growth = .03				
Implied time preference		.000	.019	.039
1953		38.7	32.1	25.8
1959		38.7	31.0	24.3
1964		38.0	29.1	21.6
c Consumption Growth = .05				
Implied time preference			.000	.019
1953			75.5	70.4
1959			83.1	76.5
1964			87.0	78.8
II. Exponential Growth Factor = 1.03				
a Consumption Growth = .00				
Implied time preference	.000	.030	.050	.070
1953	−29.3	−39.4	−42.0	−42.1
1959	−43.6	−52.5	−53.7	−52.5
1964	−60.7	−69.0	−69.1	−66.5
b Consumption Growth = .03				
Implied time preference		.000	.019	.039
1953		29.1	20.4	12.9
1959		23.6	14.1	6.5
1964		20.1	9.2	.9
c Consumption Growth = .05				
Implied time preference			.000	.019
1953			71.5	63.8
1959			73.6	64.3
1964			77.2	66.0
III. Exponential Growth Factor = 1.04				
a Consumption Growth = .00				
Implied time preference	.000	.030	.050	.070
1953	−62.2	−71.9	−72.3	−69.8
1959	−85.9	−92.7	−90.6	−85.7
1964	−114.3	−119.0	−114.6	−107.2
b Consumption Growth = .03				
Implied time preference		.000	.019	.039
1953		11.6	1.2	−6.8
1959		−.4	−11.2	−18.7
1964		−9.2	−21.3	−29.3
c Consumption Growth = .05				
Implied time preference			.000	.019
1953			59.9	49.8
1959			56.0	44.4
1964			56.9	43.3
Observed Aggregate Personal Saving[a]				
1953 61.4				
1959 72.3				
1964 91.7				

[a]See Table 1

the older population experienced. Accordingly, saving by the young will be greater (when it is positive) in order to sustain a higher level of retirement consumption than the current retirees enjoy

As in the previous set of simulations, the higher the rate of growth in consumption is relative to the rate of growth in income, the higher is the simulated value of saving. In particular, when consumption grows at a rate of 5 percent per annum and income grows at a rate of 2 or 3 percent, the simulations render apparently reasonable levels of saving given the crudeness of the simulations. However, upon further investigation, the acceptability of these values diminishes. The values of aggregate saving do not seem to differ greatly among the three years for any rate of interest. The widest range in values is exhibited in case Ic, when the interest rate is .05 Whereas observed saving increased 49 3 percent over the period 1953 64, the simulated values in case Ic imply an increase of only 15 2 percent. Extrapolating into the future (beyond 1964) or into the past (before 1953) suggests that the simulated values are unsatisfactory

D *Other Micro Foundations*

In the introduction to this paper, it was claimed that the results of this empirical investigation are strong enough to preclude using other micro foundations (utility functions) in the context of a life cycle model. That is, one might conclude from the analysis above that the utility function assumed by the authors of the life cycle is at fault However, this is not the case. Given the budget constraint, the utility function prescribes the optimal consumption path Figure 1 presents a schematic representation of the shapes of the consumption paths employed in the simulations above. Although these paths do not represent every possible consumption path, they certainly do encompass all the reasonable and many of the unreasonable paths possible. The use of a different utility function would merely change the analysis to the extent that it

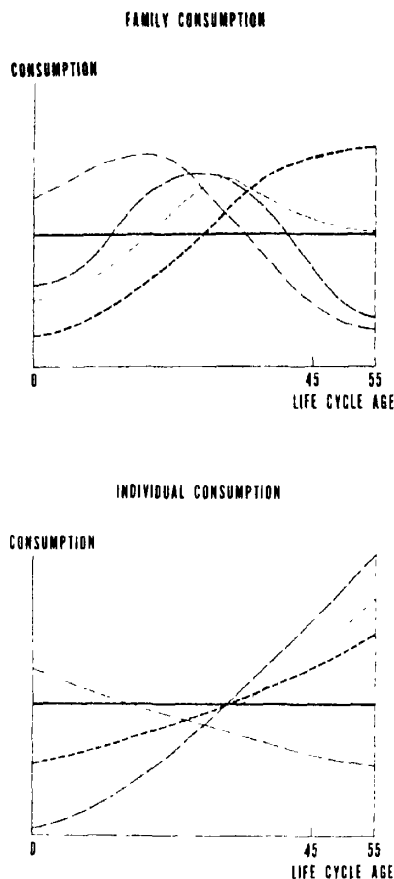



FIGURE 1 LIFETIME CONSUMPTION PATHS USED IN THE SIMULATIONS

changes the shape of the optimal consumption path. Given that such a variety of shapes are used in this analysis, the use of an alternative utility function would not change the general results of these simulations

III. Conclusion

It is now clear that the use of aggregate disposable income, which overestimates the appropriate variable (aggregate disposable labor income) and biases the simulations upward, does not vitiate the results of these simulations. On the contrary, the use of the

appropriate concept of income would merely reduce the simulated values and reinforce the conclusion drawn here.

In this paper, the life cycle model has been used to simulate values of aggregate saving. In its original form, the model is found lacking, accounting for, at best, 42 percent of observed personal saving. The inclusion of families and upward sloping lifetime income streams *reduces* the simulated values of aggregate saving. The use of age-related upward-sloping income streams decreases the simulated values if they were previously negative and increases the previously positive levels of simulated saving. Nevertheless, the model is still lacking in its ability to explain aggregate personal saving for all three years in question, and the use of a different utility function would not improve the model's explanatory ability.

One might argue that the model would render higher levels of aggregate saving if it incorporated bequests which result from uncertainty of time of death. However, the act of transferring net worth from one individual to another within the private sector does not represent an increase in personal income in the national income accounts; the transfer of assets from one individual to another does not represent an act of saving. Aggregate saving will increase as a result of bequests if the beneficiary saves a greater proportion of his unearned income than did his benefactor. However, the level of such saving is not likely to account for the large differentials between actual saving and the simulated values presented here.

One might also argue that the assumption of certainty of time of death renders the simulated values of saving artificially low. That is, because individuals do not know when they will die, they will save to provide for the contingency of living beyond the expected age of death given in life expectancy tables. It has been demonstrated elsewhere that the effects of uncertainty of time of death are not unambiguous. However, it also has been demonstrated that the life cycle model with uncertainty of time of death and a capital constraint prohibiting individuals from assuming a net debtor position is deficient in explaining aggregate personal saving. (See the author, ch. 5.)

APPENDIX

Summary of Mathematical Notation

C = aggregate consumption
Y = aggregate income
A = aggregate net worth
b_1, b_2 = regression coefficients
U = lifetime utility function
$u(c_i)$ = utility in period i derived from consuming c_i
T = years in the individual's economic life
c_i = flow of consumption services enjoyed by an individual in period i
p = rate of pure time preference
$-d$ = elasticity of marginal utility with respect to consumption
T_j = years remaining in the jth individual's economic life
y = individual's earned income
r = real rate of return
a = individual's net worth
P = population size
s = individual saving
S = aggregate saving
h = a constant
f = number of people in the household

REFERENCES

A. Ando and F. Modigliani, "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *Amer. Econ. Rev.*, Mar. 1963, *53*, 55–84.

——— **and** ———, "Tests of the Life Cycle Hypothesis of Savings: Comments and Suggestions," *Oxford Inst. Econ. Statist. Bull.*, May 1957, *19*, 99–124.

R. E. Brumberg, "An Approximation to the Aggregate Saving Function," *Econ. J.*, Mar. 1956, *66*, 66–72.

M. J. Farrell, "The Magnitude of 'Rate-of-Growth' Effects on Aggregate Savings," *Econ. J.*, Dec. 1970, *80*, 873–94.

M. R. Fisher, "Exploration in Savings Behavior," *Oxford Inst. Econ. Statist. Bull.*, Aug. 1956, *18*, 201–77.

Raymond W. Goldsmith et al., *Studies in the National Balance Sheet of the United States*, Vol. II, Princeton 1963.

J. M. Keynes, *The General Theory of Employment, Interest and Money*, Chicago 1964.

M. Landsberger, "The Life Cycle Hypothesis. A Reinterpretation and Empirical Test," *Amer. Econ. Rev.*, Mar. 1970, *60*, 175–84.

H. Lydall, "The Life Cycle in Income, Saving and Asset Ownership," *Econometrica*, Apr. 1955, *23*, 131–50.

F. Modigliani and R. Brumberg, "Utility Analysis and Aggregate Consumption Functions: An Attempt at Inegration," unpublished paper.

——— and ———, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in Harold R. Williams and John D. Huffnagle, eds., *Macroeconomic Theory: Selected Readings*, New York 1969, 99–140.

K. Nagatani, "Life Cycle Saving: Theory and Fact," *Amer. Econ. Rev.*, June 1972, *62*, 344–53.

Dorothy S. Projector, *Survey of Changes in Family Finances*, Washington 1968.

E. A. Thompson, "Intertemporal Utility Functions and the Long-Run Consumption Function," *Econometrica*, Apr. 1967, *35*, 356–61.

J. Tobin, "Life Cycle Saving and Balanced Growth," in William Fellner, ed., *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967, 231–56.

H. Watts, "Long-Run Income Expectations and Consumer Saving," in Thomas Dernburg et al., eds., *Studies in Household Economic Behavior*, New Haven 1958, 103–44.

B. B. White, "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis," unpublished doctoral dissertation, Stanford Univ. 1976.

M. E. Yaari, "On the Consumer's Lifetime Allocation Process," *Int. Econ. Rev.*, Sept. 1964, *5*, 304–17.

Board of Governors of the Federal Reserve System, Division of Research and Statistics, *Flow of Funds, Annual 1953–1964*, Washington, Nov. 6, 1976.

U.S. Bureau of the Census, *Current Population Report*, Series P-20, P-60, various issues.

———, Technical Report No. 17, *Trends in the Income of Families in the U.S., 1947–1964*, Washington 1967.

———, *Statistical Abstract of the United States 1954*, 75th ed.; *1960*, 81st ed.; *1965*, 86th ed.; *1974*, 95th ed.; Washington.

U.S. Office of Business Economics, *Surv. Curr. Bus.*, Washington 1968.

Vertical Integration: The Monopsony Case

By MARTIN K. PERRY*

If a firm behaves imperfectly in a market, then vertical integration through that market will generally alter the circumstances upon which its production choices are based. However, the only case that has been adequately discussed is that of a monopolist integrating forward into a competitive downstream industry which employs the monopolist's intermediate product in variable proportions with other inputs.[1] The primary difficulty in making a complete analysis of vertical integration by imperfectly competitive firms (including monopoly) is in defining a workable concept of forward or backward integration. In this paper, I propose a definition of backward integration which enables a substantial improvement of our insight into backward integration by imperfectly competitive buyers.[2] The case of monopsony is then analyzed in depth.[3]

The traditional measure of backward integration by a firm is the ratio of its own production of an intermediate input to its total employment of the input. Although empirically useful, such a measure defines vertical integration as the *resolution* of the firm's production decisions. Instead, I propose a measure of backward integration which is *independent* of the production decisions made by an imperfectly competitive buyer. This measure allows us to examine the *impact* of backward integration on a monopsonist's choices of input production and employment. Furthermore, partial integration becomes well-defined. Thus, *at each degree of integration* we can consider the welfare implications of and incentives for further backward integration by an imperfectly competitive buyer.

We obtain three results for the monopsony case First, further backward integration will generally, but not always, induce the monopsonist to expand its employment of the input. If integration induces an expansion of input employment, then consumers of the monopsonist's product will benefit from lower final prices. Second, further backward integration reduces the rental earnings of the remaining independent suppliers of the input. This occurs because the monopsonist reduces its purchases from these suppliers more rapidly than it integrates. Finally, for reasonable specifications of the acquisition costs of backward integration, the monopsonist has an incentive to integrate backward at least partially. However, this incentive is composed of more than the usual efficiency gains. Backward integration also enables the monopsonist to reduce the rent component of its input costs.

*Research economist, Bell Laboratories This paper is based upon my doctoral dissertation. The thesis itself was nurtured by my advisor, A Michael Spence, and readers, James N. Rosse and Bruce M. Owen This paper also benefited from comments by Dennis W Carlton, Elizabeth E Bailey, Frederick R Warren-Boulton and my colleagues at Bell Laboratories Robert D. Willig, John C Panzar, Eric B Lindenberg, and Paul S Brandon This paper represents my own views and assumptions, not necessarily those of the Bell System

[1]Monopoly pricing of the intermediate product induces inefficient substitution away from this input in the production of the final good by the downstream industry. The resulting efficiency loss creates an incentive for the monopolist to integrate forward By producing the final good itself, the monopolist can employ its input efficiently with the other inputs and thereby increase profits See the papers by Lionel McKenzie, Meyer Burstein, John Vernon and Daniel Graham, Richard Schmalensee, George Hay, and Frederick Warren-Boulton

[2]In a similar manner, we can define forward integration in order to examine such by imperfectly competitive sellers The monopoly case is reconsidered in my dissertation, ch 6

[3]Backward integration by a dominant buyer and by Cournot duopsonists has also been considered in my dissertation, chs. 4 and 5, respectively These cases are more difficult and less illustrative However, they do involve some interesting modifications.

I. The Structure of Input Supply

Assume that the monopsonist is the sole buyer of an input produced by competitive suppliers. Suppose also that the monopsonist must set a single market price for input purchases, that is, explicit price discrimination is not feasible. In general, integration by the monopsonist would be a discrete process of acquiring individual suppliers. In addition, if there were differences among the suppliers, it would also be necessary to specify which firms were acquired and which were not.[4] Thus, in order to simplify the process of backward integration, we assume instead that the competitive supply industry is characterized by "Ricardian" increasing costs, as defined by Jacob Viner, pp. 206–10. Production of the input obeys constant returns to scale throughout, but its supply price is rising because of the limited availability of a specialized factor. With this construction we can suppress the identity of input suppliers and continuously partition the industry supply into integrated and nonintegrated segments.

Assume that the variable factors are available at given prices to this industry. Let $\mathcal{C}(\mathcal{V})$ be the minimum variable cost necessary for the industry to produce $\mathcal{V}$ units of the input.[5] $\mathcal{C}(\mathcal{V})$ does not include rents to the limited factor. With no external technological economies or diseconomies, the industry's marginal variable cost $\mathcal{C}_1(\mathcal{V})$ is the supply curve for the input.[6] Because of the limited factor, marginal variable costs should rise as the industry produces more of the input. Thus, we also specify that the supply price of the input is strictly rising, i.e., $\mathcal{C}_{11}(\mathcal{V}) > 0$ for $\mathcal{V} > 0$.

We can now define the variable cost function for any subset of input suppliers. Let S be a fraction of the total quantity of the limited factor, and denote the variable cost function by $C(V,S)$. Since there are constant returns to scale in the production of the input, the cost of producing V units of the input using S of the limited factor must be the fraction S of the cost of producing V/S units of the input with all of the limited factor.

$$\text{(1a)} \quad C(V,S) \equiv S \cdot \mathcal{C}(V/S) \qquad \text{for} \quad 0 < S \leq 1$$

For $S = 1/2$, producing V with half the limited factor is half as costly as producing twice V with all of the limited factor. In other words, $C(V,S)$ is linearly homogeneous in its arguments.[7]

The required specification of $C(V,S)$ becomes more apparent upon examining the marginal cost function

$$\text{(1b)} \quad C_1(V,S) \equiv \mathcal{C}_1(V/S) \qquad \text{for} \quad 0 < S \leq 1$$

$C_1(V,S)$ is the supply curve of firms possessing S of the limited factor. It is equivalent to the industry supply curve shifted horizontally by the fraction S (see Figure 1). In other words, the industry supply curve can be continuously and horizontally partitioned into supply curves of firms owning fractions of the limited factor.[8] If there are a large number of suppliers each owning an equal fraction of the limited fac-

[4] Integration could then enable implicit price discrimination similar to the case of forward integration by a monopolist. See the author (1978).

[5] Let $\mathcal{C}(\mathcal{V})$ also be thrice continuously differentiable on $\mathcal{V} \geq 0$. Clearly, $\mathcal{C}(0) = 0$ and $\mathcal{C}_1(\mathcal{V}) > 0$ for $\mathcal{V} > 0$.

[6] Subscripts denote full or partial derivatives, while superscripts are employed for enumeration. Let $\mathcal{C}_1(0) > 0$ and $\mathcal{C}_1(\mathcal{V}) > 0$ for $\mathcal{V} > 0$.

[7] The definition of $C(V,S)$ can be illustrated for the special case where production of the input requires the limited factor N and one variable factor L. Let $\bar{N}$ be the total quantity of the limited factor and let $W = 1$ be the price of the variable factor. Assume that the production function $F(N,L)$ is thrice continuously differentiable, with positive but strictly diminishing returns to each factor, strictly declining marginal rates of substitution between the factors, and linear homogeneity. As a result, $C(V,S)$ can be implicitly defined by the expression $V = F(S \cdot \bar{N}, C(V,S))$ for $0 < S \leq 1$. Clearly, $\mathcal{C}(V) = C(V,1)$.

[8] The structure assumed on input supply is not as restrictive as it might seem. Suppose that the input supply were rising solely because of pecuniary external diseconomies (rising supply price of some factor) so that no rents were being earned. By costlessly integrating this industry, the monopsonist would be one step closer to the industry in which the rising supply

FIGURE 1

tor, then we can discuss backward integration as acquiring a fraction of the *firms* in the industry. We adopt this convention for the remainder of the exposition.

Let λ be the fraction of the input suppliers acquired by the monopsonist. For these subsidiaries, the marginal cost of internal production of the input is $C_1(V,\lambda)$ for $\lambda > 0$.[9] On the other hand, since $(1 - \lambda)$ is the fraction of the firms remaining independent, $C_1(V, 1 - \lambda)$ for $1 - \lambda > 0$ is the external supply curve of the input. I define λ as the degree of backward integration. Disintegration ($\lambda = 0$), partial integration ($0 < \lambda < 1$), and full integration ($\lambda = 1$) are now well defined, independent from the actual production and employment choices of the monopsonist.

With this structure, we can examine the monopsonist's choices of internal production, external purchase, and thus total employment of the input. These choices depend upon the degree of integration since λ determines the partition of input supply into integrated and nonintegrated segments. We can then examine the monopsonist's incentive to integrate backward. The results are derived from the properties of the variable cost function $C(V,S)$. The structure on the input supply $\mathcal{C}(\mathcal{V})$ endows $C(V,S)$ with thrice continuous differentiability and the following properties:

(2a) $C_1(V,S) > 0, C_{11}(V,S) > 0$

(2b) $C_2(V,S) < 0, C_{22}(V,S) > 0$

(2c) $C_{21}(V,S) = C_{12}(V,S) < 0$ for $V > 0$ and

(2d) $C(V,S)$ linearly homogeneous

Note that $C_1(V,S)$ and $C_2(V,S)$ are homogeneous of degree zero and that the second partial derivatives are homogeneous of degree minus one.

II. The Monopsonist's Input Employment Choice

With $C(V,S)$, we can now specify the monopsonist's expenditures in obtaining a quantity $X > 0$ of the input at each degree of integration. After doing so, we construct the monopsonist's derived demand for the input. The monopsonist's input employment at each degree of integration is then examined.

Let $X_e \geq 0$ be the quantity of the input purchased from independents, and $X_i = X - X_e \geq 0$ be the quantity of input production by the monopsonist's subsidiaries. The monopsonist's expenditures can then be expressed as

$$(3)\quad E(X,X_e,\lambda) \equiv C(X - X_e,\lambda) + X_e \cdot C_1(X_e, 1 - \lambda)$$

for $0 < \lambda < 1$

The monopsonist will choose its level of input purchases so as to minimize expenditures at each level of input employment and each degree of integration. This requires the monopsonist to equate the marginal expenditure from the two input sources:

---

price was due to a limited factor earning rents. The best example of a limited factor is land which is fixed in quantity and specialized either by composition or climate to the production of a particular crop. Other limited factors might include natural resources such as mineral ores or human resources such as talent.

[9]This construction rules out economies or diseconomies of backward integration.

$$(4)\quad C_1(X - X_e, \lambda) = C_1(X_e, 1 - \lambda) + X_e \cdot C_{11}(X_e, 1 - \lambda)$$

for $\quad 0 < \lambda < 1 \quad$ and $\quad X > 0$

Note that the monopsonist's marginal expenditure on input purchases exceeds the marginal production cost by the increment in rents on inframarginal purchases. Assuming the second-order condition is satisfied, expression (4) defines the input purchases function $X_e(X, \lambda)$; it is continuously differentiable on $X \geq 0$ and $0 \leq \lambda \leq 1$.[10]

The monopsonist's *minimized* expenditure function can now be defined solely in terms of the quantity of input employed and the degree of integration:

$$(5a)\quad E(X, 0) \equiv X \cdot C_1(X, 1)$$

$$(5b)\quad E(X, \lambda) \equiv C(X - X_e(X, \lambda), \lambda) + X_e(X, \lambda) \cdot C_1(X_e(X, \lambda), 1 - \lambda)$$

for $\quad 0 \leq \lambda < 1$, and

$$(5c)\quad E(X, 1) \equiv C(X, 1)$$

If the monopsonist is disintegrated, $\lambda = 0$, its expenditures are entirely purchase costs, whereas if the monopsonist is fully integrated, $\lambda = 1$, its expenditures are entirely production costs. However, if the monopsonist is partially integrated, $0 < \lambda < 1$, then expenditure minimization implies that both types of costs will be incurred, i.e., $0 < X_e(X, \lambda) < X$.[11]

Expenditure minimization by the partially integrated monopsonist implies an inefficient production of the input. This is a consequence of monopsony behavior with respect to *only* the input supply by independents. Input production per firm is higher for the monopsonist's subsidiaries than for independent suppliers. This fact is easily derived from the expenditure minimization condition. Using the facts that $C_{11}(V, S) > 0$ and $C_1(V, S)$ is homogeneous of degree zero, (4) implies that

$$(6)\quad \frac{X - X_e(X, \lambda)}{\lambda} > \frac{X_e(X, \lambda)}{1 - \lambda} \quad \text{for } 0 < \lambda < 1$$

The monopsonist's subsidiaries are utilized more intensively than independents. This inefficiency in the production of the input under partial integration is borne jointly by the monopsonist (lower profits) and the independents (lower rents). Thus, it is conceptually distinct from our primary concern with the inefficient *employment* of the input which adversely affects consumers.

Having specified the monopsonist's expenditures in obtaining the input, we need only characterize the monopsonist's revenue from employing the input in order to determine the profit-maximizing level of input employment at each degree of integration. Let $NR(X)$ be the *net* revenue attributable to the monopsonist's employment of X of the input. This is the revenue from sales of the output produced with X of this input and the profit-maximizing levels of other inputs (obtainable at given prices) *minus* the expenditures on these other inputs. Assume $NR(X)$ is a well-defined, twice continuously differentiable function on $X \geq 0$. The marginal net revenue function $MNR(X) \geq 0$ is the monopsonist's inverse derived demand for the input. As a result of decreasing returns to scale in the production of the *final* good and/or downward-sloping *final* demand, assume that the inverse derived demand function is nonincreasing i.e., $MNR_1(X) \leq 0$.[12]

Assuming that the monopsonist is initially disintegrated, the rents earned by the input suppliers prevent the monopsonist from integrating costlessly. Define $A(\lambda)$ as the acquisition cost to the monopsonist of obtaining λ of the suppliers. These costs are previously agreed upon payments to the former suppliers. For this reason they are

[10] Note that $X_e(0, \lambda) = 0$, $X_e(X, 0) = X$, and $X_e(X, 1) = 0$. For the second-order condition to be violated, it is necessary but not sufficient that the marginal expenditure on input purchases be a *declining* function of the level of those purchases over some range (see my dissertation, p. 30).

[11] See my dissertation Appendix 2.2.

[12] There are no special restrictions on the production function of the final good or the state of final market competition. We require only that the monopsonist's inverse derived demand function be continuously differentiable and nonincreasing.

independent of the monopsonist's current input purchase and employment choices. Thus, these choices can be examined without specifying $A(\lambda)$.

The monopsonist's profit function can now be expressed as

$$(7)\quad \pi(X,\lambda) \equiv NR(X) - E(X,\lambda) - A(\lambda)$$

The monopsonist will choose its level of input employment so as to maximize profits at each degree of integration. This requires the monopsonist to equate the marginal revenue with the marginal expenditure:[13]

$$(8)\qquad MNR(X) = E_1(X,\lambda)$$

Assume the second-order conditions for a nontrivial solution are satisfied.[14] If so, expression (8) defines the input employment function $X(\lambda)$; it is continuously differentiable on $0 \leq \lambda \leq 1$.

The input employment function $X(\lambda)$ has important welfare implications. Integration has the effect of reducing the monopsonistic restriction on the employment of the input. As such, integration results in an increased employment of the input and a larger output of the final good. The monopsonist's consumers would then benefit from lower final prices.

Clearly, full integration eliminates the inefficient underemployment of the input. The monopsonist's expenditures to obtain an additional unit of the input are reduced to just the marginal production costs of that unit. Input employment is expanded, i.e., $X(1) > X(0)$.[15] This outcome duplicates the market equilibrium when competitive buyers constitute the derived demand for the input.

Despite the fact that $X(1) > X(0)$, it need not be the case that *further* backward integration will always induce the monopsonist to expand input employment. However, Proposition 1 demonstrates that $dX(\lambda)/d\lambda > 0$ should be considered the "normal" case.

PROPOSITION 1: *For $0 \leq \lambda < 1$, the input employment of the monopsonist will expand, remain constant, or contract with further integration as the marginal expenditure function is respectively rising, unchanging, or declining at the profit-maximizing choice of input employment. That is,*

$$(9)\quad sgn\left[\frac{dX(\lambda)}{d\lambda}\right] = sgn\,[E_{11}(X(\lambda),\lambda)] \qquad for \quad 0 \leq \lambda < 1$$

PROOF:[16]

Differentiating (8) with respect to λ yields

$$(10)\quad \frac{dX(\lambda)}{d\lambda} = \frac{E_{12}(X(\lambda),\lambda)}{MNR_1(X(\lambda)) - E_{11}(X(\lambda),\lambda)}$$

The denominator of this expression is negative by the second-order condition for profit maximization. Partially differentiating $E_1(X,\lambda)$ (from fn. 13) with respect to λ and employing Euler's theorem on $C_1(V,S)$ and $C_{11}(V,S)$, we find that

$$(11)\quad E_{12}(X,\lambda) = \left[\frac{\partial X_e(X,\lambda)}{\partial\lambda} + \frac{X_e(X,\lambda)}{1-\lambda}\right] \cdot [2\cdot C_{11}(X_e(X,\lambda), 1-\lambda) + X_e(X,\lambda) \cdot C_{111}(X_e(X,\lambda), 1-\lambda)] \quad \text{for} \quad 0 \leq \lambda < 1$$

After partially differentiating (4) with respect to λ and again employing Euler's theorem on $C_1(V,S)$ and $C_{11}(V,S)$, we find that inequality (6) implies that the first bracketed term of (11) is negative. Furthermore, it can easily be shown that $E_{11}(X,\lambda)$ has the same sign as the second bracketed term of (11). Thus, $E_{12}(X,\lambda)$ has the opposite sign of $E_{11}(X,\lambda)$. Therefore, $dX(\lambda)/d\lambda$ has the same sign as $E_{11}(X(\lambda),\lambda)$ for $0 < \lambda < 1$.

[13] Using (5b) and simplifying via (4), note that $E_1(X,\lambda) = C_1(X_e(X,\lambda), 1-\lambda) + X_e(X,\lambda) \cdot C_{11}(X_e(X,\lambda), 1-\lambda)$ for $0 \leq \lambda < 1$.

[14] For the second-order condition to be violated, it is again necessary but not sufficient that the marginal expenditure function on input purchases be declining over some range (see my dissertation, p. 45).

[15] Since $C_{11}(X,1) > 0$, we find that $E_1(X,1) = C_1(X,1) < C_1(X,1) + X \cdot C_{11}(X,1) = E_1(X,0)$ for all $X > 0$. Thus, (8) and the corresponding second-order conditions imply that $X(1) > X(0)$ (see Figure 2). Also see S. Y. Wu, pp. 118–19, for a graphical exposition of this result.

[16] The details of this proof are contained in my dissertation Appendices 2.5 and 2.6.

Limiting arguments complete the proof for $\lambda = 0$.

The case of $E_{11}(X(\lambda),\lambda) > 0$ with $dX(\lambda)/d\lambda > 0$ should be considered the "normal" case. If $E_{11}(X(\lambda),\lambda) < 0$, the marginal expenditure on input purchases from independent suppliers must be *nonincreasing* with the level of purchases, that is, the second bracketed term of (11) must be negative when evaluated at $X(\lambda)$. Such cases cannot occur if $C_{111}(X_e(X(\lambda),\lambda), 1-\lambda) > 0$ and need not occur even if $C_{111}(X_e(X(\lambda),\lambda), 1-\lambda) < 0$. The example of linear marginal costs ($C_{111}(V,S) = 0$) provides an illustration of a "normal" case for Proposition 1. Since $E_{12}(X,\lambda) < 0$ in the linear example, backward integration by the monopsonist shifts its marginal expenditure curve outward from $E_1(X,0)$ toward the industry supply curve $E_1(X,1)$ (see Figure 2). The monopsonist's profit-maximizing input employment increases, resulting in lower final prices for consumers.

III. The Earnings of Independent Suppliers

Having determined the input employment function $X(\lambda)$, we can now characterize the rents accruing to independents at each degree of integration. The impact of integration on the earnings of independents can then be examined. In addition, the acquisition cost function $A(\lambda)$ can be specified, thereby enabling an examination of the monopsonist's incentive to integrate backward.

FIGURE 2

The total rents accruing to independent suppliers are the receipts from sales of $X_e((X(\lambda),\lambda)$ at the corresponding supply price minus the variable costs of production. The rents per fractional unit of the total suppliers are

$$(12)\quad r(\lambda) \equiv \left[\frac{1}{1-\lambda}\right]\cdot[X_e(X(\lambda),\lambda) \cdot C_1(X_e(X(\lambda),\lambda), 1-\lambda) - C(X_e(X(\lambda),\lambda), 1-\lambda)] \quad \text{for} \quad 0 \le \lambda < 1$$

Proposition 2 characterizes the behavior of $r(\lambda)$.

PROPOSITION 2: *Declining (constant) marginal net revenue implies that per unit rents decline (remain constant) as the monopsonist integrates backward. That is,*

$$(13)\quad \frac{dr(\lambda)}{d\lambda}\begin{Bmatrix} < 0 \\ = 0\end{Bmatrix} \quad \textit{if} \quad MNR_1(X(\lambda))\begin{Bmatrix} < 0 \\ = 0\end{Bmatrix} \quad \textit{for} \quad 0 \le \lambda < 1$$

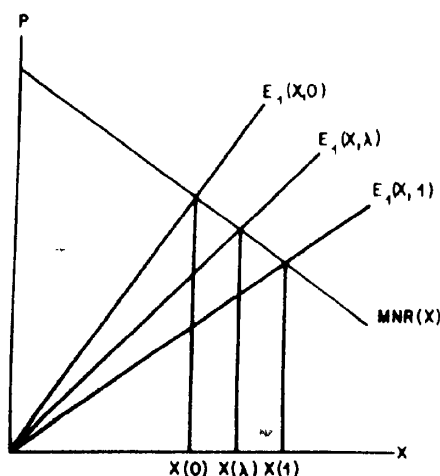
PROOF:[17]

Euler's theorem on $C(V,S)$ implies that

$$(14)\quad r(\lambda) = -C_2(X_e(X(\lambda),\lambda), 1-\lambda) \quad \text{for} \quad 0 \le \lambda < 1$$

Differentiating $r(\lambda)$ with respect to λ and employing Euler's theorem on $C_2(V,S)$, we find that

$$(15)\quad \frac{dr(\lambda)}{d\lambda} = -C_{21}(X_e(X(\lambda),\lambda), 1-\lambda) \cdot \left[\frac{dX_e(X(\lambda),\lambda)}{d\lambda} + \frac{X_e(X(\lambda),\lambda)}{1-\lambda}\right]$$

After totally differentiating (8) with respect to λ (using $E_1(X(\lambda),\lambda)$ from fn. 13), employing Euler's theorem on $C_1(V,S)$ and $C_{11}(V,S)$, and substituting from (10) and

[17] The details of this proof are contained in my dissertation Appendix 2.7.

(11), we find that the bracketed term of (15) has the same sign as $MNR_1(X(\lambda))$. Since the first term of (15) is positive, $dr(\lambda)/d\lambda$ also has the same sign as $MNR_1(X(\lambda))$.

Proposition 2 states that those suppliers remaining independent will incur lower earnings as the monopsonist integrates backward. The integrating monopsonist reduces its input purchases more rapidly than the reduction in the independent supply of the input. This fact can be shown by rearranging the bracketed term of (15) which is negative when $MNR_1(X(\lambda)) < 0$:

$$(16)\qquad -\frac{dX_e(X(\lambda),\lambda)}{X_e(X(\lambda),\lambda)} > -\frac{d(1-\lambda)}{1-\lambda}$$

As a result, the price received by independent suppliers, $C_1(X_e(X(\lambda),\lambda), 1-\lambda)$, declines as the monopsonist integrates backward.[18] Thus, independent suppliers would oppose backward integration by the monopsonist.

IV. The Monopsonist's Incentive to Integrate Backward

To examine the monopsonist's incentive to integrate backward, we must first specify the acquisition costs $A(\lambda)$. These costs are related to the rents accruing to suppliers before they are acquired. Although any acquisition cost function could be specified, only two will be examined. The first specification allows the monopsonist to acquire any fraction of the total suppliers *only* by paying them the per period *initial* rents per fractional unit:

$$(17)\qquad A^1(\lambda) \equiv \lambda \cdot r(0)$$

The second specification allows the monopsonist repeatedly to acquire *only* a small additional fraction of the total suppliers at the *prevailing* rents per fractional unit:

$$(18)\qquad A^2(\lambda) \equiv \int_0^{\lambda} r(\xi)\cdot d\xi$$

Under $A^1(\lambda)$, the monopsonist is not allowed to acquire identical suppliers at different acquisition prices. Under $A^2(\lambda)$, the monopsonist adjusts its profit-maximizing purchases and input employment after each small increase in its degree of integration and can then acquire a small additional fraction of the total suppliers at the new unit rents. The specification of $A^2(\lambda)$ collapses this story into a instantaneous process. By Proposition 2 (for $MNR_1(X) < 0$), the total and marginal acquisition costs are lower under $A^2(\lambda)$ than under $A^1(\lambda)$, i.e., $A^2(\lambda) < A^1(\lambda)$ and $r(\lambda) < r(0)$ for all $\lambda > 0$. Thus, $A^2(\lambda)$ would encourage a greater degree of integration by the monopsonist.

Neither $A^1(\lambda)$ or $A^2(\lambda)$ involves "threats" by the monopsonist against independent suppliers. If refusal-to-deal threats were viewed as credible by independents, then the acquisition cost function would take the form $A^3(\lambda) \equiv 0$. Similarly, the acquisition cost function

$$A^4(\lambda) \equiv \lambda \cdot \lim_{S\to 1} r(S)$$

requires the credible threat against each supplier that he will be the only firm not to have sold out to the monopsonist. Finally, the acquisition cost function $A^5(\lambda) \equiv \lambda \cdot r(\lambda)$ might be reasonable if independents were convinced of the monopsonist's intention to acquire λ of their ranks. Each supplier would then be indifferent between selling out and remaining independent. Unlike $A^3(\lambda)$, $A^4(\lambda)$, and $A^5(\lambda)$, both $A^1(\lambda)$ and $A^2(\lambda)$ imply that the acquired suppliers were never paid less than they were earning at the point of acquisition. *In this sense*, the monopsonist's integration via $A^1(\lambda)$ or $A^2(\lambda)$ cannot be considered "predatory" even though per unit rents earned by the remaining independent suppliers have declined.

The monopsonist's incentive to integrate backward is defined as the change in maximized profits which result from an increase in the degree of integration λ. Totally differentiating $\pi(X(\lambda),\lambda)$ with respect to λ and taking note of condition (8), the incentives to integrate under $A^1(\lambda)$ and $A^2(\lambda)$ are:

$$(19)\qquad \frac{d\pi^1(\lambda)}{d\lambda} = -E_2(X(\lambda),\lambda) - r(0)$$

[18] See my dissertation Appendix 2.8

$$(20) \quad \frac{d\pi^2(\lambda)}{d\lambda} = -E_2(X(\lambda), \lambda) - r(\lambda)$$

$E_2(X(\lambda), \lambda)$ is the reduction in minimized expenditures, required to obtain the profit-maximizing level of input employment, as a result of a marginal increase in the monopsonist's degree of integration.[19] These savings must be weighed against the acquisition costs of a marginal increase in the degree of integration, $r(0)$ for $A^1(\lambda)$ and $r(\lambda)$ for $A^2(\lambda)$. These forces can be unambiguously resolved for the case of $\pi^2(\lambda)$.

PROPOSITION 3: *At each degree of integration λ, the monopsonist facing acquisition costs of $A^2(\lambda)$ has a further incentive to integrate backward. That is,*

$$(21) \quad \frac{d\pi^2(\lambda)}{d\lambda} > 0 \quad \textit{for all} \quad 0 \le \lambda < 1$$

PROOF:[20]

Using Euler's theorem on $C(V,S)$ and $C_1(V,S)$, $d\pi^2(\lambda)/d\lambda$ can be expressed solely in terms of $C(V,S)$, $C_1(V,S)$, and $C_{11}(V,S)$. Simplifying by means of (4) and employing the linear homogeneity of $C(V,S)$ and zero homogeneity of $C_1(V,S)$, we find that

$$(22) \quad \frac{d\pi^2(\lambda)}{d\lambda} = \left[\frac{X(\lambda) - X_e(X(\lambda), \lambda)}{\lambda} - \frac{X_e(X(\lambda), \lambda)}{1-\lambda}\right] \cdot C_1\left(\frac{X(\lambda) - X_e(X(\lambda), \lambda)}{\lambda}, 1\right) - \left[C\left(\frac{X(\lambda) - X_e(X(\lambda), \lambda)}{\lambda}, 1\right) - C\left(\frac{X_e(X(\lambda), \lambda)}{1-\lambda}, 1\right)\right]$$

for $\quad 0 < \lambda < 1$

By (2a) and (6), the first term is positive and the second term is negative. Since $C_{11}(V,S) > 0$, the mean value theorem implies that $d\pi^2(\lambda)/d\lambda > 0$ for $0 < \lambda < 1$. Limiting arguments complete the proof for $\lambda = 0$.

The monopsonistic restriction of input purchases results in a supply price for independent suppliers and corresponding rental earnings which understate their value to the monopsonist. In addition, the independent suppliers are *not* undervalued to potential outside investors since these investors could not alter the restriction of input purchases. Thus, the monopsonist will initiate backward integration. If the partially integrated monopsonist need pay only the prevailing unit rents $r(\lambda)$ to integrate further, then this argument applies at each degree of integration and the monopsonist will choose to integrate fully.

Proposition 3 also provides information about the incentive to integrate for other acquisition cost functions. Acquisition cost functions $A(\lambda) \le A^2(\lambda)$ for all λ will also induce the monopsonist to integrate fully. In other words, acquisition strategies which involve "credible threats" against independents, for example, $A^3(\lambda)$, $A^4(\lambda)$, or $A^5(\lambda)$, would certainly imply full integration by the monopsonist. On the other hand, acquisition cost functions $A(\lambda) > A^2(\lambda)$ for all λ may not induce the monopsonist to integrate fully. The function $A^1(\lambda)$ is a member of this latter class. Since $d\pi^1(0)/d\lambda = d\pi^2(0)/d\lambda > 0$, there exists an initial incentive to integrate under $A^1(\lambda)$. However, further integration may well be unprofitable beyond some degree of less than full integration, i.e., $\max\ \pi^1(\lambda) = \pi^1(\lambda^1)$ where $\lambda^1 < 1$.[21]

V. Interpretation of the Incentive to Integrate

As a result of Proposition 3, the monopsonist's incentive to integrate backward cannot be interpreted as strictly an internali-

[19] Using (5b) and simplifying via (4), note that $E_2(X, \lambda) = C_2(X - X_e(X, \lambda), \lambda) - \lambda_e(X, \lambda) \cdot C_{12}(X_e(X, \lambda), 1 - \lambda)$ for $0 < \lambda < 1$.

[20] The details of this proof are contained in my dissertation Appendix 2.9.

[21] If $F(N, L) = A \cdot N^\alpha \cdot L^{1-\alpha}$ for $0 < \alpha < 1$ and $MNR(X) = b \cdot X^{-1/\eta}$ over a meaningful range $(X \ge \epsilon > 0)$, then λ^1 can be calculated as a function of the parameters α and η. See my dissertation pp. 74–75. For example, if $\alpha = \eta = 1/2$ then $\lambda^1 \approx$ 68. In addition, λ^1 increases with η and α.

zation of efficiency losses. To demonstrate this fact, suppose that the marginal net revenue curve is perfectly inelastic over the relevant range, i.e., $X(\lambda) = X(0)$ for all $0 \leq \lambda \leq 1$. No efficiency loss from underemployment of the input would exist for the monopsonist to profitably internalize by integrating. However, the proof of Proposition 3 remains valid for this case. The monopsonist will fully integrate under $A^2(\lambda)$ and partially integrate under $A^1(\lambda)$. This result can be explained by what I shall call the "rent effect." Define *total* rent payments as rents earned by independents $((1 - \lambda) \cdot r(\lambda))$ *plus* the acquisition costs $(A(\lambda))$. *Backward integration enables the monopsonist to reduce these total payments.* Define $\mathcal{R}(\lambda)$ as this reduction in rents:

$$(23) \quad \mathcal{R}(\lambda) \equiv r(0) - [(1 - \lambda) \cdot r(\lambda) + A(\lambda)]$$

Under either $A^1(\lambda)$ or $A^2(\lambda)$, the monopsonist never pays more than the initial unit rents in acquiring suppliers. Under $A^2(\lambda)$, total rents are continually reduced by the acquisition process, implying that the rent reduction increases with the degree of integration, i.e.,

$$(24) \quad \frac{d\mathcal{R}^2(\lambda)}{d\lambda} = -(1 - \lambda) \cdot \frac{dr(\lambda)}{d\lambda} > 0$$

for $MNR(X)$ less than perfectly elastic. Thus, the monopsonist will integrate fully even when there is no efficiency loss to internalize.

However, under $A^1(\lambda)$ the reduction in total rents can result only from the reduced rents earned by the remaining independents, as is evident from

$$(25) \quad \mathcal{R}^1(\lambda) = (1 - \lambda) \cdot [r(0) - r(\lambda)]$$

Thus, when there exists no efficiency loss to internalize, the profit-maximizing degree of integration λ^1 must be less than unity since $\mathcal{R}^1(\lambda) > \mathcal{R}^1(0) = \mathcal{R}^1(1) = 0$ for $0 < \lambda < 1$. The reduction in rents paid to independents more than offsets the monopsonist's share of the efficiency loss in input *production* as a result of partial integration, $0 < \lambda^1 < 1$. Thus, $\pi^1(\lambda^1) > \pi^1(0) = \pi^1(1)$. Finally, the increase in profits from the reduction in total rents under $A^1(\lambda)$ or $A^2(\lambda)$ need not be trivial.[22]

For strictly downward-sloping marginal net revenue functions, the monopsonist's incentive to integrate results from both the rent effect and an "efficiency effect." The efficiency effect is the increase in profits to the monopsonist which results from the internalization of the efficiency loss from under-*employment* of the input. As the monopsonist integrates, the expansion of input employment converts this efficiency loss into profits. A naive extrapolation from the monopoly case would portray the efficiency effect as *the* incentive to integrate. But by Proposition 2, the monopsonist's incentive to integrate backward can result *purely* from the efficiency effect *only if* the marginal net revenue function is perfectly elastic ($MNR_1(X) = 0$) preventing unit rents from declining with integration. Thus, the general incentive of a monopsonist to integrate backward must be understood in terms of the rent effect as well as the efficiency effect.

VI. Conclusion

In summary, complete backward integration eliminates the efficiency losses from monopsonistic behavior. The expansion of input employment by the integrated monopsonist results in greater final output at lower prices for consumers. The efficient outcome of a market with a competitive derived demand as well as a competitive supply would be duplicated. With the acquisition cost function $A^2(\lambda)$, the profit-maximizing monopsonist chooses to integrate fully. Only the original input suppliers are injured by the monopsonist's integration. However, since their reduced rental earnings are a direct transfer of income to the monopsonist, they could be compensated so as also to benefit. The acquisition cost function $A^1(\lambda)$ would fully compensate the original fixed-factor owners *if* the monopsonist chose to integrate fully. But as we illustrated, the

[22] If $X(\lambda) = X(0)$ for $0 \leq \lambda \leq 1$ and $F(N, L) = A \cdot N^{1/2} \cdot L^{1/2}$, then the percentage reduction in total rents will be 50 percent under $A^2(\lambda)$. See my dissertation, pp. 79–82.

monopsonist may well choose not to integrate fully under $A^1(\lambda)$. Efficiency losses (in the employment of the input and in the production of the input) would remain, consumer benefits would be curtailed, and the remaining independent suppliers would still experience reduced rental earnings

In conclusion, vertical integration by imperfectly competitive firms can be successfully modeled and analyzed if the concept of vertical integration, full and partial, is well defined and meaningful. Integration requires the acquisition of firms in competitive stages of the industry. Partial integration leaves the imperfectly competitive firm with a market vs. nonmarket choice, the resolution of which influences its further incentives to integrate. Not only can these incentives to integrate be specified and examined but also the welfare implications of integration on all participants in the industry can be considered.

REFERENCES

M. L. Burstein, "A Theory of Full-Line Forcing," *Northwestern Univ Law Rev.*, Feb. 1960, *55*, 62 95.

G. A. Hay, "An Economic Analysis of Vertical Integration," *Ind. Org. Rev*, 1973, *1*, 188 98.

L. McKenzie, "Ideal Output and the Interdependence of Firms," *Econ. J.*, Dec. 1951, *61*, 785 803.

M. K. Perry, "The Theory of Vertical Integration by Imperfectly Competitive Firms," unpublished doctoral dissertation, Stanford Univ., Center Res. Econ. Growth, res. memo. series, no. 197, Dec. 1975.

——, "Price Discrimination and Foward Integration," *Bell J. Econ.*, Spring 1978, *9*, 209 17.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 442-49.

J. M. Vernon and D. A. Graham, "Profitability of Monopolization by Vertical Integration," *J. Polit. Econ.*, July/Aug. 1971, *79*, 924 25.

J. Viner, "Cost Curves and Supply Curves," *Z Nationalokon.*, Sept. 1931, *3*, 23 46; reprinted in Kenneth E. Boulding and George J. Stigler, eds., *Readings in Price Theory*, Homewood 1952.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, Aug./Sept. 1974, *82*, 783-802.

S. Y. Wu, "The Effects of Vertical Integration on Price and Output," *Western Econ J*, Spring 1964, *2*, 117-33

Market Behavior with Demand Uncertainty and Price Inflexibility

By DENNIS W. CARLTON*

Most economists would agree that the large majority of markets do not precisely fit the classical assumptions of competition. For many markets, prices do not adjust at each instant of the day to balance supply and demand. Moreover, firms often do not know how much of their product will be demanded each day.

There are good reasons why most markets depart from the strict classical assumptions (see Armen Alchian). Changing prices frequently is time consuming and costly. Consumers may dislike price fluctuations More importantly, prices may have to remain in effect for some time if their "signal" is to be received. The demand that an individual firm sees is random because the number of customers that frequent the firm will generally vary from day to day. In formulating its operating policy, a firm must take into account the randomness of its demand. Firms do not feel that they can sell all they want at the current market price and are concerned with overproducing or having excess capacity. Firms are also concerned with underproducing or having too little capacity. In these markets, it is an outcome of the market process that occasionally some customers will be unable to purchase the good instantly.

For these uncertain markets, the amount that a firm is willing to supply depends not only on the current market price, but also on the entire stochastic structure of demand that it faces. In this environment, supply cannot be defined without first specifying the random structure of demand.

There will be three essential features of market operation that we will study; price inflexibility, demand uncertainty, and timing considerations. By price inflexibility, I do not mean that prices do not respond to permanent shifts in the underlying supply and demand factors, but only that prices cannot be adjusting at each instant of time. An important feature of the analysis will be to determine exactly how prices are endogenously determined by market forces. Demand uncertainty means that, at the beginning of any market period after prices have been set, firms do not know for sure what their demand will be, although they do know what the random distribution of demand looks like. Demand is uncertain over the period for which prices are inflexible. Timing considerations refer to the need to have produced or to have made some prior commitment to production, such as the purchase of equipment, before the unknown customer demand is observed.

It is not immediately clear what the consequences of these three nonclassical features of market operation are, even though these three features would appear to be realistic characterizations of many market operations. In this paper, I address the following questions: How do firms compete in such markets? Can equilibrium be meaningfully defined, and if so, how does it compare to the classical equilibrium when the uncertainty is removed from the demand side? What are the properties of the competitive equilibrium as the size of the market increases? Will this equilibrium be Pareto optimal? In this equilibrium, do firms produce too little of the good in question? Would society benefit if the government paid lump sum subsidies to firms so as to encourage them to expand their production of the good?

For the markets under study, it will be a natural feature to have some customers being unable to purchase the good, and

*Department of economics, University of Chicago I wish to thank Gary Becker, Franklin Fisher, Peter Diamond, Robert Hall, Lester Telser, and Sam Peltzman for valuable advice I thank George Borts and the anonymous referees for their helpful comments

some firms being unable to sell all of their stock, or equivalently use all their capacity. Each good will have two characteristics associated with it, namely its price and the probability that it can be purchased. Customers will have preferences not only for the price of the good, but also for the probability of obtaining it. Firms will compete amongst themselves until an equilibrium is reached. Market clearing will require equilibrium along the dimensions of both price and probability of obtaining the good. In equilibrium, supply will not in general equal demand, and there will always be some customers who are unable to purchase the good. The "customers" can also be interpreted as being other firms who are trying to buy factor inputs for their production process. With this interpretation, we obtain a model where it is perfectly natural for firms to be concerned with obtaining an "assured" supply of the input, a concern that appears uppermost in the minds of businessmen (see Alfred Chandler).

In the special case where social welfare is measured by expected surplus, we show that a competitive equilibrium is optimal. This result stands in sharp contrast to previous models in the literature on optimal pricing under uncertainty. However, in general, a competitive equilibrium will not lead to the socially optimal point. The social optimum will, under a plausible set of assumptions, involve paying lump sum subsidies to encourage firms to expand their production.

The model is applicable to any market where availability of the good or of the means to produce the good is important. It does appear that in the private sector for many industries demand fluctuations are not always absorbed by price changes and that changes in the probability of whether or when the good can be obtained is often an important equilibrating mechanism. Some examples include retail stores, hotels, restaurants, and manufacturing. In the regulated and government sector the model also seems to have wide-spread application. For example, for airlines, railroads, public parks, and electric utilities, prices do not vary continuously and uncertainty in demand is absorbed by changes in rationing frequency.

I. Competitive Market Clearing

There is a large literature on the effects of uncertainty on firm behavior.[1] Analyses of competitive markets focus on the effect of having uncertainty in price, and they maintain the assumption that firms can always sell all they want at the future uncertain market price.[2] There are never any shortages in equilibrium. In his pioneering works, Edwin Mills (1959, 1962) has examined the effect of demand uncertainty and price inflexibility on the behavior of a monopolist who must decide what price to charge and how much to produce before demand can be observed. Surprisingly, despite the realism of the assumptions of demand uncertainty, price inflexibility, and a lead time necessary for production, there has been no attempt to examine the implication of these assumptions within a competitive environment.[3] The purpose of this paper is to provide such an examination, and to derive and investigate the properties of an equilibrium in which it is natural to have supply not equal to demand. A simple model is presented which attempts to capture the essential features of the markets under study

There are N identical firms who compete with each other. To make the assumption of competition plausible, the number of firms N will be considered to be large enough to prevent firms from having any monopoly power. Firms maximize expected profits.

At the beginning of each period, each firm sets price, which remains in effect for the entire period, and it decides how much of the good to produce and stock for the period.[4] No additional production or de-

[1] See Michael Rothschild and John McCall and the references cited therein.

[2] See, for example, Edward Zabel

[3] Since this paper was written, John Gould and Arthur DeVany and Thomas Saving have investigated issues closely related to those of this paper.

[4] It is not necessary for the good to be produced at the very beginning of the market period. All that is required is that some commitment to production, such

livery of the good can occur during the period. The production cost per unit of the good is c, where c is strictly positive. To keep the model simple, it is assumed that the good is perishable so that it is impossible to hold inventories between periods. Provided holding inventory is a costly activity, the qualitative results derived below will be unchanged.

I now wish to generate a simple characterization of random demand per firm in which the number of customers that a firm sees is random. One way to generate such a random process would be to have three random states of nature for each consumer: a) one in which the consumer does not desire the good (the occurence of this state depends on exogenous random variables); b) one in which the consumer desires to purchase some amount of the good but cannot; and c) one in which the consumer desires to purchase some amount of the good and does so. (For simplicity, let each consumer have the same per capita demand curve for the good when the good is desired.) The utility of the consumer in each of the three states would influence the consumer's total expected utility. Conditional on not being in state a), the amount that firms produce will influence whether a consumer winds up in state b) or c). Since it is this last set of events that we wish to analyze, I suppress discussion of state a) for the remainder of the paper (except for footnotes), not because it is unimportant but because its inclusion, though straightforward, would be cumbersome, and would obscure the main thrust of the analysis. It is crucial to stress that the main ideas and conclusions of the paper apply to any demand process that generates random demands per firm.

I adopt the following simple process to generate random demand per firm Suppose that each period there are L identical consumers who desire to purchase the good, each with per capita demand $x(p)$. During each period, each customer randomly frequents a firm.[5] The customer knows the price the firm charges, and through reputation, the probability that the good will be available at the firm. If a customer finds a firm out of the good, he simply leaves the firm and does not obtain the good for that period. He does not search at the other firms.[6] Buyers have preferences for not only how much they purchase and spend on the good, but also for the probability of being able to buy the good. Therefore, competition does not force firms to necessarily charge the same price but rather to offer price-shortage combinations which leave the customer at the same level of expected utility.

Equilibrium in an uncertain market is said to exist when 1) consumers are indifferent at which of the firms they shop each period, and 2) no firm behaving optimally can offer a price-shortage combination which would leave consumers better off, and which would allow the firm to earn nonnegative expected profits.[7] Before examining how market equilibrium is determined, let us first look at the incentives facing individual consumers and firms.

II. Consumer Behavior

In calculating his expected utility from going to any firm, a customer is concerned

as the purchase of equipment or other inputs, be made prior to observing demand

[5] In equilibrium the demand to a firm is a random variable from a binomial process of size L and probability $1/N$ For large N, which is assumed, this binomial process converges to the Poisson which in turn converges to the normal for moderate size L. (It is this normal approximation that is used in the Appendix, Section A) The reader can think of the demand arrival process as either Poisson or equivalently (since the number of firms can be finite but large) as binomial with large N (See William Feller, ch 7) As stressed above, what is of crucial importance is that demand to the individual firm be random. Whether total industry demand is random (as is natural with a Poisson interpretation) or nonrandom (as is natural with a binomial interpretation) is irrelevant The referee has noted that nothing in the model requires L or N to be finite, only that L/N be finite

[6] Just as in the case of inventory holding, consumer search behavior, providing it is costly, would not alter the main features of the model This point is discussed more fully below

[7] This equilibrium concept is closely related to the equilibrium concept in hedonic markets See Sherwin Rosen

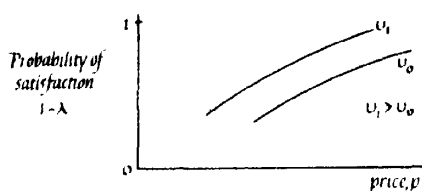



FIGURE 1. ISO-UTILITY CONTOURS

with both the probability $1 - \lambda$ of obtaining the good and the price p charged for the good. We can write this expected utility as $U(1 - \lambda, p)$ The function U defines the iso-utility contours between $1 - \lambda$ and p that leave a consumer indifferent.[8] Typical iso-utility contours are drawn in Figure 1

The diagram shows that along any iso-utility curve, as price rises, the probability of satisfaction must rise if consumers are to remain indifferent. Also, for any fixed probability of satisfaction, consumers always prefer lower prices.

Consumers will always try to reach their highest iso-utility contour, and will only go to a firm that they think will provide this highest iso-utility level. If the buyer believes that several firms provide this highest utility level, then he will choose among them randomly.

No strong conclusion about the shape of the iso-utility curves are justified. We impose the very weak assumptions that the iso-utility curves exist over the relevant range[9] in $(1 - \lambda, p)$ space, that they are continuous, and that they satisfy an upper and lower Lipschitz condition. This latter condition postulates that there exist two numbers, b and B, such that $0 < b < B < \infty$ and such that the slope along any iso-utility curve always lies between them. The Lipschitz requirements insure that the consumer is never willing to make infinite tradeoffs in either the p or $1 - \lambda$ directions.

III. Behavior of the Firm

Since consumers will wind up going only to those firms that provide the highest utility level in the market, competition forces firms to take the utility level as given. (If instantaneous production were possible so that no shortages could occur, then each good would have only one characteristic (price) associated with it. In that case, utility-taking behavior is equivalent to price-taking behavior and this market would behave exactly as a classical supply and demand analysis would indicate.) At the beginning of each period, firms have to decide on a price and production policy so as to maximize their profits subject to the constraint that they provide at least the given level of utility to consumers. Firms know that if they remain competitive with the other firms, then they will receive their random share of demand.

We can write the total amount that the firm decides to produce at price p as $s \cdot x(p)$. The variable s can be interpreted as the maximum number of customers that a firm can satisfy in that period. Henceforth, we will refer to s as customer capacity. Clearly, the amount that a firm decides to produce affects the probability that a customer will be able to obtain the good from that firm.

Let us examine the relation between the expected number of customers M, who will

[8] The formula for an iso-utility contour is easy to derive Consider a two-good world Let $u(x_1, x_2)$ be the person's (von Neumann-Morgenstern) utility function Good 1 is subject to shortages, while good 2 is always available at price of 1 Let $V(y, p)$ be the indirect (von Neumann-Morgenstern) utility function when income is y and price of good 1 is p Then expected utility is $E(U) = (1 - \lambda)V(y, p) + \lambda u(0, y)$ When good 1 is available at price p, the consumer purchases $x(p)$ units of it where, by Roy's identity, $x(p) = -V_p/V_y$ where subscripts denote partial differentiation. If I adopted the more general formulation of random demand where I allow for the possibility that consumers might not desire the good in some periods, I would obtain the expression for total expected utility by multiplying the previous expression for $E(U)$ by the (exogenous) probability of desiring the good and then adding to this expression the expected utility of the consumer when he does not desire the good.

[9] By this assumption I simply mean that there is some range of prices, which includes $p = c$, the cost of production, for which the consumer is interested in purchasing the good. In other words, if the consumer does not have positive demand for prices near c, then the market for the good will not exist, and there is nothing to analyze.

find the firm out of the good, and the customer capacity s that the firm provides. Let $pr(i)$ stand for the probability that i customers will arrive at the firm. Then, we can write that

$$(1) \qquad M(s) = \sum_{s+1}^{\infty} (i - s)pr(i)$$

If all N firms follow the same operating policies, then the total expected number of customers who will be dissatisfied is $N \cdot M$, and the fraction of dissatisfied customers will equal NM/L. The fraction $1 - \lambda$ of customers who are able to obtain the good can be written as

$$(2) \qquad 1 - \lambda(s) = 1 - \frac{N \cdot M(s)}{L}$$

(In my 1975 paper I explain how (2) can be interpreted as applying to a firm even when all firms do not follow the same policies.) In the Appendix, Section A, I show that using the normal distribution to approximate the discrete binomial process of customer arrival, the probability of satisfaction function, $1 - \lambda(s)$ can be written as

$$(3) \qquad 1 - \lambda(s) = \frac{\sigma I(u) + s}{\sigma^2}$$

where $\sigma^2 = L/N$,

$$I(u) = \int_{-\infty}^{u} [t - u] f(u)\, du$$

$f(u)$ is the normal density function, and $u = (s - \sigma^2)/\sigma$.

For a given level of utility, firms want to choose a price p and a customer capacity s, so that profits are maximized and the consumer is able to achieve the given level of utility. When firms remain competitive by offering the given level of utility, they randomly receive their equal share of the L customers. Letting $pr(i)$ stand once again for the probability that i of the L customers visit a firm this period, we can write that expected profits equal

$$(4) \qquad \pi(s,p) = p \cdot x(p) \sum_{0}^{s} i\, pr(i) + px(p)s \sum_{s+1}^{\infty} pr(i) - csx(p)$$

The first term in (4) represents expected sales revenue when $i \leq s$ customers come to the firm, while the second term represents expected sales revenue when more than s customers come to the firm. The last term in (4) is the cost of being able to service s customers. Since (3) expresses a one-to-one relation between the probability of satisfaction $1 - \lambda$ and the customer capacity s, we can interpret (4) as expressing profits as a function of $1 - \lambda$ and p.

Regarding profits as a function of $1 - \lambda$ and p, we can draw iso-profit curves in $(1 - \lambda, p)$ space. A typical family of such curves is depicted in Figure 2. The two iso-profit curves at the far right of the diagram are drawn to illustrate that each iso-profit curve involving positive profits can "turn around" on itself as price rises sufficiently high to drive demand to zero. Since consumers always prefer to be on the northwest boundary of the iso-profit curves, competition will insure that the "dotted" segments of the iso-profit curves are never observed. The heavy dotted line in the diagram represents the π_s curve which is derived by setting $\partial\pi/\partial s = 0$ in (4). Iso-profit curves cross the π_s curve vertically, and so the relevant portions of all iso-profit curves emanate from the π_s curve. For the relevant portions of the curves, with fixed price, profits decrease as probability of satisfaction increases. Hence, in the diagram $\pi_1 < \pi_2$. The curve on the far left of Figure 2 represents the zero-profit ($\pi = 0$) curve, which is the only one this paper will be interested in.

For any given iso-utility level, $\bar{u}$, the firm will choose to operate at a point of tangency between the iso-utility curve representing iso-utility level, $\bar{u}$, and the highest

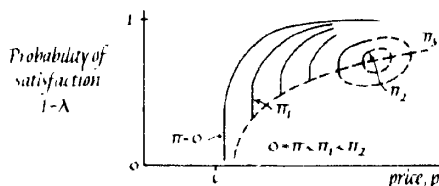



FIGURE 2. ISO-PROFIT CURVES

iso-profit curve. No firm ever chooses to operate to the left of the $\pi = 0$ curve since that represents negative expected profits.

The properties of the zero-profit curve play a key role in determining the behavior of equilibrium as the number of customers per firm increases. I now describe the relevant properties[10] of the zero-profit curve, proved in a separate appendix available on request from the author.

The $\pi = 0$ curve is concave (i.e., $d^2(1 - \lambda)/dp^2 < 0$), starts off with a very large slope at a point a little to the right of $p = c$ on the horizontal axis, rises to 1 as price increases, and has a very flat slope for sufficiently high prices. The curve always lies to the right of the vertical line $p = c$, since price must cover not only production costs, but also the cost of unsold goods. As price rises, firms can afford to provide a larger customer capacity s. Hence along the $\pi = 0$ curve the probability of satisfaction increases to 1 as price increases.

As the customer per firm ratio L/N increases, the $\pi = 0$ curve is affected in several ways. First, the entire curve shifts up, indicating that for fixed price as the number of customers per firm increases firms can afford to increase their customer capacity in such a way that there is a higher probability of satisfying customers. Basically, this result occurs because there are economies of scale in servicing a stochastic market. The proportional risk of having unsold goods declines as the customer per firm ratio increases. In other words, to achieve a satisfaction probability of .5 in a market with 100 customers per firm requires a $s/100$ figure that is larger than the $s/1,000$ figure in a market with 1,000 customers per firm.[11] As the customer per firm ratio continues to increase, the $\pi = 0$ curve shifts up to the $1 - \lambda = 1$ line.

How does the slope of the $\pi(1 - \lambda, p) = 0$ curve behave as the customer per firm ratio increases? For any fixed price p greater than c, the slope $(d(1 - \lambda)/dp)$ falls monotonically to zero as L/N increases. Furthermore, for any fixed probability of satisfaction, $1 - \lambda$, below 1, the slope $d(1 - \lambda)/dp$ approaches infinity as L/N increases.

IV. Market Equilibrium

In the diagram of the iso-profit curves, superimpose the iso-utility curves of customers. We can define a contract curve as the locus of tangencies between the iso-utility and iso-profit curves. Firms always operate on this contract curve.

In a classical market, firms compete with each other by offering to consumers lower prices (i.e., higher utilities) than other firms. Prices (or consumer utilities) continue falling (rising) until firms have no incentive to lower price (raise utility) any more. Analogously, in this market, firms compete with each other by offering better (i.e., higher utility) combinations of price and probability of satisfaction to consumers. The utility level is "bid" up until there is no incentive for any firm to continue to alter its price-probability of satisfaction combination. This point occurs when the contract curve intersects the zero-profit ($\pi = 0$) curve. At this point, firms would prefer to go out of business rather than offer a higher utility combination to consumers and earn negative expected profits. Hence, competition on the utility level forces the market equilibrium up the contract curve, until the zero-profit curve is reached. Equilibrium can be regarded as a tangency[12,13] between the zero-profit curve and the highest at-

[10]Since I am using a continuous random variable to approximate a discrete positive random variable, there is a slight error involved. By the Central Limit Theorem, we know that any such approximation errors become insignificant for even moderate (i.e., 15–20) values of the customer per firm ratio, L/N. In the subsequent analysis, I ignore such approximation errors.

[11]Recall that s refers to customer capacity.

[12]Note the similarity between this equilibrium and the hedonic good equilibrium of Rosen.

[13]I ignore the uninteresting case of corner solutions in which either a) the firms produce nothing or b) each firm by itself stocks an amount of the good to satisfy the entire customer population *by itself*. Notice that multiple tangencies are possible, since the iso-utility curves can be convex. The dynamic arguments justifying the establishment of this equilibrium are outlined in the author (1975). See the author (1977a) for a discussion of possible instabilities in these markets.

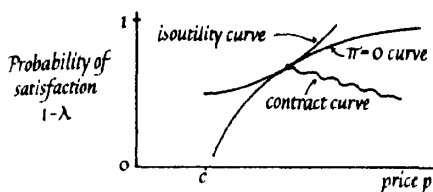



FIGURE 3 MARKET EQUILIBRIUM

tainable iso-utility curve.[14] This equilibrium is depicted in Figure 3.

There are several noteworthy features of this equilibrium. In general, there will be a positive probability of being unable to obtain the good. Second, in equilibrium the price will exceed the constant cost of production. This occurs because the revenue from sold goods must compensate not only for the cost of producing those goods, but also for the cost of producing the unsold goods. Equivalently, it is necessary to pay for available but unused capacity. Third, there is no reason why supply should equal demand even in expected values since equilibrium depends in part on consumers' willingness to take risk.

V. Behavior of Market Equilibrium as the Customer Per Firm Ratio Increases

Armed with the properties of the $\pi = 0$ curve, we can investigate the behavior of equilibrium as L/N, the customer per firm ratio, increases. It will be useful for the reader to recall that b and B are the lower and upper bounds on the slope of the iso-utility curves, respectively.

THEOREM 1: *As L/N, the customer per firm ratio, increases, the equilibrium price associated with the market-clearing point approaches the deterministic market-clearing price c.*

[14]With instantaneous production, the model becomes identical to the classical supply and demand model For that case, the $\pi = 0$ curve is a vertical line at $p = c$, and equilibrium as defined above coincides with the classical equilibrium of price $= c$, probability of satisfaction $= 1$. We see then that the classical model is a special case of this model.

PROOF:

The method of proof will be to show that as L/N increases, any equilibrium point $(p^*, 1 - \lambda^*)$ of the market clearing under uncertainty will eventually lie to the left of the vertical line $p = c + e$ for every positive e.

Choose the point $p = c + e$ for any positive e. Equilibrium in the uncertain market is defined as a point of tangency between the $\pi = 0$ curve and an iso-utility curve. Now, increase L/N. As L/N increases, the slope of the $\pi = 0$ curve declines to zero for any fixed $p > c$. Increase L/N so that the slope of the $\pi = 0$ curve is less than b at $p = c + e$. This implies that the slope of $\pi = 0$ is less than b for all $p \geq c + e$ since the $\pi = 0$ curve is concave. But then it is impossible for any iso-utility curve to be tangent to the $\pi = 0$ curve at any price above $c + e$. Hence, any market equilibrium price p^* is less than $c + e$. Since p^* must be greater than c for any production to occur at all, and since p^* is less than $c + e$ for any positive e, it follows that

$$\lim_{L/N \to \infty} p^* \to c$$

THEOREM 2: *As L/N, the customer per firm ratio, increases, the equilibrium probability of satisfaction approaches 1.*

PROOF.

The method of proof will be to show that as L/N increases, any equilibrium point $(p^*, 1 - \lambda^*)$ lies above the horizontal line defined by probability of satisfaction $= 1 - \bar{\lambda}$ for $1 - \bar{\lambda} < 1$.

As before, equilibrium is determined by a point of tangency between the $\pi = 0$ curve and an iso-utility curve. Choose any $1 - \bar{\lambda} < 1$. Increase L/N. As L/N increases, the slope along $\pi = 0$ curve at the point associated with a probability of satisfaction equal to $1 - \bar{\lambda}$ becomes arbitrarily large. Continue increasing L/N until the slope at $1 - \bar{\lambda}$ on the $\pi = 0$ curve exceeds B. Because of the concavity of the $\pi = 0$ curve, the slope along the $\pi = 0$ curve exceeds B for all $1 - \lambda < 1 - \bar{\lambda}$. Hence, for suffi-

ciently large L/N it is impossible for any iso-utility curve to be tangent to the $\pi = 0$ curve for a probability of satisfaction less than or equal to $1 - \bar{\lambda}$. Since the equilibrium probability of satisfaction is bounded above by 1, and lies above every $1 - \bar{\lambda}$ less than 1, it follows that

$$\lim_{L/N \to \infty} 1 - \lambda^* \to 1$$

It immediately follows from Theorems 1 and 2 that the equilibrium level of expected utility achievable by consumers in equilibrium approaches the level of utility achievable in the deterministic market, where price equals c and the probability of satisfaction equals 1.

THEOREM 3: *As L/N, the customer per firm ratio, increases, the percent discrepancy between the amount supplied and the amount demanded approaches* 0.

PROOF:

The total amount demanded equals the number of customers times the per capita demand $L \cdot x(p)$, while the total amount supplied equals the number of firms times the customer capacity per firm times the per capita demand $N \cdot s \cdot x(p)$ To prove the theorem it is sufficient to show that $N \cdot s/L \to 1$ as L/N increases.

Write the zero-profit condition as

$$(1 - \lambda)p \cdot L = Nc \cdot s \tag{5}$$

From the previous two theorems we know that in equilibrium $p \to c$ and $1 - \lambda \to 1$ as L/N increases. Hence the theorem follows immediately.

Theorem 3 dealt with the percent discrepancy between supply and demand. What about the absolute discrepancy, $[L - N \cdot s]x(p)$ does that too approach zero as the customer per firm ratio L/N increases? The answer in general is no. Usually the absolute discrepancy will approach either plus or minus infinity as L/N increases. In other words, equilibrium is possible even though the number of dissatisfied customers is arbitrarily large.

The reason why the market equilibrium does not converge to the deterministic one in all respects as the customer per firm ratio L/N increases can be explained as follows. As L/N increases, the total uncertainty in the market increases, so that market operation under uncertainty differs considerably from that under certainty. On the other hand, by the law of large numbers, the proportional risks caused by the uncertainty vanish as L/N increases. Therefore, percentage-wise concepts (for example, supply ÷ demand), or concepts that apply to individual units of the good (for example, price) or individual customers (for example, probability of satisfaction) approach their values in the corresponding deterministic market as L/N increases. However, aggregate concepts such as supply, demand, and total number of customers dissatisfied do not in general approach their values in the deterministic market as the customer per firm ratio increases.[15,16]

VI. Different Types of Customers

It is perfectly natural to imagine a market with two types of customers who have different preferences between price and

[15] How much do markets under uncertainty differ from those under certainty? As seen below, social welfare implications and incentives for vertical integration are different See the author (1977a) for further differences It is possible (see the author, 1975) to calculate lower bounds on the customer per firm ratio necessary to achieve any given level of convergence of price and probability to the certainty equilibrium $p = c$, $1 - \lambda = 1$ For convergence to the 1 percent level, L/N must exceed 6,500

[16] The number of firms N and the total amount demanded at any price p, $Lx(p)$ have been taken as fixed The assumptions were made for analytical tractability It is not necessary that total demand be fixed For example, total demand could be random and each firm could obtain some fixed share of total demand As long as demand to an individual firm is random, the analysis developed above applies See Section I for further discussion Are there any incentives for merger in the model? When total demand is random and firms obtain fixed shares of total demand, no incentives for merger exist When total demand is fixed, it appears that total merger is desirable But this last result emerges only because considerations like spatial patterns of demand variation and costs of merger do not explicitly appear in the model

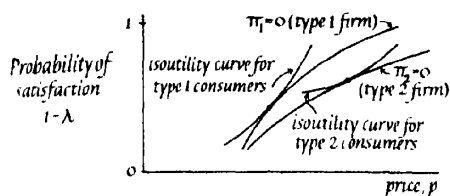



FIGURE 4. SPECIALIZED EQUILIBRIUM

FIGURE 5. NONEXISTENCE OF SPECIALIZED EQUILIBRIUM

probability of satisfaction. In such a situation, it is possible to have an equilibrium in which two types of firms are established, each of which caters only to the preferences of one type of consumer. An equilibrium involving firm specialization is depicted in Figure 4.[17]

As Figure 5 illustrates, such specialized equilibrium may not always exist.[18] The specialized equilibrium cannot exist because all the type 2 customers are better off at type 1 stores than at type 2 stores.

When only one equilibrium can exist in the market, the question of where it is established will be determined by the tastes of the majority. If any firm does not cater to the tastes of the majority, it will lose a majority of its business and will have to specialize in the minority's tastes. But, by assumption, specialized equilibrium is impossible, so the firm could not profitably attract just the minority types to its firm.

VII. Search Behavior

So far, the model has restricted consumer's search to only one firm per period. Unlike other models of search, in this model consumers have information about the characteristics (i.e., $1-\lambda, p$) of each firm. Consumers simply do not know which firms have the good. If we allow consumers more than one search, then the main features of market behavior are unchanged in the sense that firms and consumers will still take into account the probability they cannot sell or buy a good. A new feature that does emerge is that all firms may not behave identically. Some firms could charge low prices and run out frequently, while others could charge high prices but run out infrequently. Customers might go to the low-price firm first and then, if unable to satisfy their demands, go to the high-priced firm. Price distributions arise naturally. Notice, though, that the higher prices compensate firms for the higher risk of having unsold capacity. The different price firms sell different products.[19]

VIII. Social Welfare Implications

The previous section examined how markets operate when the production decision must be made before the uncertain demand for the product can be observed, and when prices, once set, cannot vary over the market period. An important question is whether a competitive equilibrium involves a combination of price and probability of satisfaction that is optimal in the sense of maximizing some measure of social welfare.

[17] For the case of equilibrium involving firm specialization an outside observer might incorrectly conclude that there was a distribution of prices for an identical good and attribute it to consumer ignorance. As this paper emphasizes, since each type of firm offers a different probability of satisfaction, the "goods" at different types of firms are not identical.

[18] The nonexistence of such equilibria occurs for reasons similar to those studied by Rothschild and Joseph Stiglitz. See their article for further discussion. As that discussion makes clear, the above analysis of specialized equilibrium is different from that of a hedonic market (see Rosen). The key difference is that in the above model one of the characteristics of the good is endogenous and cannot be specified independently of consumer behavior as is true in a hedonic market.

[19] As the referee has pointed out, if consumers had imperfect information about price and availability, then just as in the other search models (see Gerard Butters), firms offering the same product (i.e., availability) could receive different prices.

This is the issue that is examined in the next two sections.

Throughout this examination, I do not allow insurance markets to develop. There are well-known reasons why such markets may not exist. For example, in the present case, there would be the problem of ascertaining that someone actually attempted to purchase the good. Such insurance markets rarely exist in practice. (If a customer finds that a firm is out of a good, there is not a market to compensate him.)

The first question to be asked is when, if ever, will the market equilibrium maximize the expected value of the total surplus to society. This question is motivated by two considerations. First, deterministic markets in competition maximize surplus, so it is natural to see if uncertain markets do also. Second, expected surplus is often used as a measure of social welfare.[20] I will show that, in the special case where expected consumer's surplus is derived from an individual's preferences between price and probability of satisfaction, a competitive equilibrium does indeed maximize the expected value of surplus to society.

Consumer's surplus is an appropriate measure of welfare only under very narrow assumptions. Moreover, in an uncertain setting expected consumer's surplus may not properly reflect consumer attitudes toward risk. To avoid the defects associated with consumer surplus, I also examine the social welfare question in a simple two-good model. A two-good model is set up by introducing an alternative (nonrationed) good, and asking how a social planner who takes both markets into account would operate this economy so as to maximize the expected utility of a representative consumer. It will be shown that the socially optimal solution will usually be different from a competitive equilibrium. The socially optimal solution will, under a plausible assumption, involve paying lump sum subsidies to the firms that deal with the good that is subject to shortages. Compared to the social optimum, a competitive equilibrium will usually not devote sufficient resources to production of the good that is subject to shortages.

[20]See, for example, Gardiner Brown and M. Bruce Johnson.

IX. Maximizing Expected Surplus

As mentioned above, consumer surplus is not generally a good measure of social welfare for uncertain markets because, aside from problems associated with its use in a deterministic setting, it may not reflect consumer preferences between the probability of obtaining the good and the price of the good. For the special case[21] where expected consumer surplus does reflect consumer attitudes toward risk, we want to examine whether a competitive equilibrium maximizes expected surplus. The main result of this section is that for this special case a competitive equilibrium does indeed maximize the expected surplus to society.

The model is the same as before. Let us consider the expression for expected surplus to society when all N firms follow the same price and stocking policy. Expected surplus to society (ESS) equals the per capita consumer surplus times the number of customers times the expected fraction of customers that are satisfied minus the cost of the goods. Expressed mathematically, we have that

$$ESS = (1 - \lambda(s)) \int_0^{x(p)} x^{-1}(q)\, dq \cdot L - c\, s\, N\, x(p) \tag{6}$$

where, to refresh the reader's mind, we repeat the definitions:

L = number of (identical) customers in the market

[21]This special case occurs when, in the notation of fn. 8, $u(x_1, x_2) = g(x_1) + x_2$, where u is the von Neumann-Morgenstern utility function, x_1 is the amount of the good that is subject to shortages, and x_2 represents all other goods which are assumed to be available at a price of one. The above analysis also applies to the more general formulation of random demand in which consumers have the possibility of not desiring the good (see the discussion in Section I about state of nature a)) provided utility in that state of nature is equal to x_2.

N = number of firms
$x(p)$ = the per capita demand curve[22]
$x^{-1}(q)$ = the inverse per capita demand function
p = price of the good
c = cost per unit of the good
s = the number of customers that can be serviced per firm
$1-\lambda$ = the probability of satisfaction as a function of s

The government wishes to determine an operating policy (i.e., s and p), so that ESS is maximized when all firms behave according to this operating policy. To maximize ESS with respect to p and s, take derivatives of (6) to obtain the following first-order conditions:

$$(1-\lambda)x'(p)pL - csNx'(p) = 0$$

or equivalently

$$(7) \qquad (1-\lambda)pL - csN = 0$$

and

$$(8) \qquad \frac{d(1-\lambda)}{ds} \cdot L \int_0^{x(p)} x^{-1}(q)\,dq - cx(p)N = 0$$

Equations (7) and (8) determine the s and p of the operating policy for each firm that the government should follow to maximize the expected total surplus to society.[23] Using the expression for profits derived earlier, it can be seen that (7) is equivalent to the condition that expected profits per firm equal zero. Equation (8) determines the point along the zero-profit ($\pi = 0$) curve at which the government should operate.

The question then arises as to whether a competitive market equilibrium would maximize the expected value of surplus to society if consumers' tradeoffs between the price of the good and the probability of obtaining the good were adequately represented by the expected value of their consumer surplus. At first glance, the answer to this question appears obvious. If expected consumer surplus reflects consumer preferences, then we know from the properties of equilibrium in an uncertain market that the expected consumer surplus of each individual (ESI) is maximized. Hence, the social planner will maximize the surplus to society at this point. This reasoning is faulty although the conclusion turns out to be correct. The sum of individual consumer surpluses does not equal the surplus to society for the markets under study. There are unsold goods at the end of each period which must enter the government's calculation of surplus but not that of any individual.

More specifically, suppose each of the L consumers seeks to maximize[24]

$$(9) \quad ESI = (1-\lambda)\int_0^{x(p)} x^{-1}(q)\,dq - px(p)$$

where the notation was just defined beneath (6). Summing ESI over all L consumers and comparing this sum to the objective function ESS of the government, we see that the two expressions differ by

$$(1-\lambda)pLx(p) - csx(p)N = x(p)[p(1-\lambda)L - sNc]$$

This last expression is the difference between the expected revenue to be received and the cost of all the goods, sold and unsold. In view of the differences in the objective functions between the individual and the government, it is interesting that the following theorem holds.

THEOREM 4: *Suppose ESI, as defined in (9), represents consumer preferences between the price p and probability of satisfaction $1-\lambda$. Then, a competitive equilibrium maximizes the expected value of surplus to society, ESS.*

[22]Recall from fn. 8 that $x(p)$ is simply the per capita demand for the good when it is available. The relation between $x(p)$, and $U(1-\lambda, p)$ is explained in that same footnote. We always assume $x'(p) < 0$.

[23]As before I disregard the uninteresting case of boundary solutions and assume that (7) and (8) have a solution that represents an interior maximum (i.e., second-order conditions for a maximum are fulfilled).

[24]Recall that consumers will maximize ESI if their von Neumann-Morgenstern utility functions are of the form $u(x_1, x_2) = g(x_1) + x_2$ where x_1 = the good under analysis, and x_2 = all other goods always available at a price of 1.

PROOF:

If ESI reflects consumer preferences, then from the definition of competitive equilibrium, we know that competitive equilibrium occurs at that point along the zero-profit curve that maximizes ESI. From (7), we know that the point that maximizes ESS also occurs along the zero-profit curve.

The difference between surplus to society ESS, and the sum of consumer surpluses to an individual $L \cdot ESI$, was derived above and equals $x(p)[(1 - \lambda)pL - Nsc]$. However, from (7), we see that along the zero-profit curve, this difference equals zero. Therefore, along the zero-profit curve, the two measures ESS and $L \cdot ESI$ attain their maximum values at the same points.

We see then that if individual consumer preferences are represented by ESI, then just as in deterministic markets, a competitive equilibrium will maximize the expected value of the total ESS. Notice that price exceeds c and firms earn zero-expected profits when expected surplus is maximized.[25] These results contrast sharply with those of other models that appear in the public finance literature (see Brown and Johnson; Michael Visscher), and deal with a similar type of problem.[26] The results of those other models imply that to maximize expected surplus to society, price should in general be less than or equal to c, and hence firms should operate at an expected loss.

The reason for the difference in results stems from the manner in which the randomness is introduced into the demand curve and the way goods are rationed. In the model under study, a firm's demand is multiplicative and equals $x(p) \cdot i$ where $x(p)$ equals per capita demand and i equals the random number of consumers who visit the firm. All customers face the same probability of being rationed. In Brown and Johnson, rationing is done by willingness to pay with the demand for units that generate large consumer surplus being satisfied first. As Visscher points out, it is difficult to imagine how such a rationing scheme could be implemented without using a recontracting market. In Visscher's models, more realistic rationing schemes are introduced, however only the case of additive demand uncertainty is analyzed. For most purposes, the multiplicative formulation would appear more plausible.[27] See the author (1977b) for further discussion.

If ESI does not represent consumer preferences for the probability of satisfaction $1 - \lambda$, and the price p, then Theorem 4 will not hold. However, if ESI does not represent consumers' preferences toward risk, then expected surplus is a very poor criterion to use as a measure of market performance in an uncertain environment.[28] In the next section, I allow the consumer to have quite general preferences for the probability of satisfaction and the price, and examine how the introduction of an alternative good affects the analysis of the social optimum.

X. The Social Optimum in a Simple Two-Good Model

Let there be two goods on which each of the L consumers can spend their identical endowment Y. Good 1 is subject to shortages, while good 2 is always available from the outside world at a constant price. The price of good 1 is p, while the price of good 2 is one. As before, each unit of good 1 uses up c units of resources and must be produced before any firm observes its random demand. Demand is random in the same manner as discussed previously. As usual, no firm can receive delivery of the good

[25] It should be mentioned that the result that expected profits equal zero at the social optimum does not depend on the particular random process for demand. See the author (1977b).

[26] To see the relation of the model of this paper to the peak load problem under uncertainty, reinterpret c as the fixed cost per unit. In the model the marginal variable cost β is taken as 0. If β is nonzero, prices would rise by β. The model above suggests that under certain assumptions $p > c + \beta$ is optimal. The previous models in the literature suggest that $p \leq c + \beta$ is optimal.

[27] This is one reason why econometric equations are specified so often in *log-log* form.

[28] This point is not addressed by either Brown and Johnson or Visscher.

once a market period has begun. The government owns each of the N firms that dispense good 1, and wishes to choose the same tax policy and operating policy for each of the N firms so as to maximize the expected utility of a representative consumer. The government faces the budget constraint that the sum of the firms' expected profits plus the total taxes collected or dispersed must equal zero.

Let $u(x,z)$ represent the von Neumann-Morgenstern utility function of each consumer where x denotes good 1 and z denotes good 2. When good 1 is obtainable at price p, the utility of each consumer is given by $V(p, Y)$, the indirect utility function. When good 1 is not obtainable, the utility of each consumer is given by $u(0, Y)$. If $1 - \lambda$ represents the probability of obtaining good 1, then the expected utility of a representative consumer can be written as

$$U(1 - \lambda, p) = (1 - \lambda)V(p, Y) + u(0, Y)$$

The government seeks to determine a transfer T for each individual,[29] a price p, and a customer capacity s (recall that s refers to the maximum number of customers that can be serviced at any firm in any market period), so that the expected utility of each (identical) consumer is maximized. The government's problem can be written as

$$(10)\quad \max_{s,p,T}\ (1 - \lambda(s))V(p, Y + T) + \lambda(s)\ u(0, Y + T)$$

subject to the budget constraint,

$$(11)\qquad \pi(s,p) - (L/N)T = 0$$

where $\pi(s,p)$ is the expression for expected profits per firm, which can be written as

$$(12)\quad \pi(s,p) = (1 - \lambda(s))p\frac{L}{N}x(p) - csx(p)$$

where $1 - \lambda(s)$ is the expression for the probability of satisfaction as a function of customer capacity s, and is given by (3).

From the statement of the problem, we see that if (and only if) the transfer T equals 0 in the socially optimal solution, then it follows that a competitive equilibrium will also be the socially optimal point since both points maximize expected utility subject to the constraint that expected profits are zero. In general, there is no reason to expect that the optimal solution to the above problem will have $T = 0$, so that a competitive equilibrium will usually not represent the social optimum. The social optimum will usually involve either taxes or subsidies for the firms who sell good 1, the good subject to shortages. In such cases government intervention into a competitive market may be called for.

In order to investigate the conditions under which either taxes or subsidies will be paid in the social optimum, it is necessary to make an assumption about consumers' preferences.

Assumption 1: The marginal utility of an extra dollar, when good 1 is obtainable, is higher than the corresponding marginal utility when good 1 is unobtainable. More precisely, $V_2(p, Y) > u_2(0, Y)$ for all p, Y, where the subscripts denote partial differentiation.

The assumption reflects the idea that the greater the variety of goods that can be purchased, the higher is the marginal utility of an extra dollar. (One sufficient condition for this assumption is that $u_{21} \geq 0$.) Given the above assumption, the following theorem holds.

THEOREM 5: *Under Assumption 1 and the assumption that per capita demand depends positively on income, the social optimum involves operating the N firms that sell good 1 at a loss and using lump sum taxes to subsidize their operation.*

PROOF:

See the Appendix, Section B.

Under Assumption 1, the socially optimal solution involves operating the N firms that produce good 1 at a loss, and

[29]The variable T is the transfer from the firms to each consumer. Hence, if $T < 0$, consumers pay a tax while firms receive a subsidy.

using lump sum taxes to subsidize these firms' revenues. Since a single price competitive equilibrium involves zero profits, we see that government intervention into a competitive market may be necessary to achieve the social optimum.[30]

The heuristic reason why, under Assumption 1, it is optimal to tax consumers and pay subsidies to firms can be seen as follows. There are two states in which the consumer can wind up, one where he can purchase the good at the market price and one where he cannot. Under Assumption 1, the last dollar is more valuable in the state in which the good is obtainable than in the state in which the good is unobtainable. A person could increase his utility if he could in some way transfer part of his income between the two possible states. Such transfers of income are impossible in the problem under examination. (Remember no insurance or recontracting markets exist.) However, what is possible is that the government can use taxes to reduce the income of a consumer in both states and subsidize the operation of firms that produce the good, and thereby reduce the price of the good subject to shortages. In this way, a transfer of purchasing power can occur between the two possible states in which the consumer can find himself. It turns out that this price reduction is always sufficient to overwhelm the decline in income, so that under Assumption 1 imposing some taxes always raises expected consumer utility.

Theorem 5 tells us that a competitive equilibrium will not achieve the social optimum. Can we say whether, under Assumption 1, a competitive equilibrium will devote too few resources to the production of good 1 and/or will involve a higher price for good 1 than occurs in the social optimum? Without further restrictions, all that can be said is that in the social optimum either the probability of satisfaction (or equivalently the customer capacity, s) will be higher and/or the price of good 1 will be lower than in a competitive equilibrium. We expect the normal case to involve an increase in the probability of satisfaction $1 - \lambda$, and a decrease in the price p. For this normal case it immediately follows that under Assumption 1 a competitive equilibrium will involve devoting too few resources (i.e., $csNx(p)$) to the production of the good that is subject to shortages, when compared to the social optimum.

XI. Summary and Conclusions

This paper has examined the behavior of markets characterized by price inflexibility, demand uncertainty, and production lags. It appears that many markets in the private, regulated, and government sectors are better described by the model examined here than by the traditional supply and demand model. Examples of markets described by the models of this paper include transportation, manufacturing, retailing, electric utilities, restaurants, hotels, and public parks.

Natural features of the markets studied here are that buyers and sellers will always have some probability of being unable to buy or sell all they want of the good. Supply will not in general equal demand. Price will exceed average production cost of the goods that are sold. It was possible to prove that as the size of the market increased, the equilibrium under uncertainty converged percentage-wise to that under certainty. Numerical calculations suggested that the customer per firm ratio might have to be unrealistically large before close convergence occurred. The social welfare implications of markets under uncertainty differ from those under certainty In the special

[30]As should be clear from the proof of the theorem, if we replace Assumption 1 with the (less plausible) assumption that the marginal utility of income declines as the variety of goods increases, then in the social optimum firms would be taxed and consumers subsidized Under the assumption that $V_2(p, Y) = u_2(0, Y)$ (as in Section IX) the optimal tax is zero (see the Appendix, Section B) Competitive equilibrium is optimal in this special case. In the more general formulation of random demand in which the consumer has the possibility of not desiring the good (see the discussion in Section I about state of nature a)), there is another term in the expression for $E(U)$ for this additional state of nature. An assumption that the marginal utility of income is lower in this state of nature than in those where the consumer desires the good is then sufficient for Theorem 5 to hold

case where expected consumer surplus reflects consumer's preferences, a competitive equilibrium is optimal. This result contrasts with previous results in the literature. In general though, the competitive equilibrium is not socially optimal, and the conditions under which subsidization would occur were derived.

The models examined here are rich in their implications. Demand uncertainty imposes costs on a firm in the form of potential idle capacity. Firms have an incentive to stabilize their random demand by finding "loyal" customers. Such loyal customers get a price discount because they enable a firm to better plan its production. In terms of the model, any customer willing to order in advance need pay only c for the product. Customers not willing to order in advance pay a higher price for the privilege of only buying occasionally. Newspaper subscriptions illustrate this point nicely. Coupons in packages are another example where random buyers pay a higher price than repeat buyers. Special favors and discounts for steady customers provide a final important example.

When the customers are interpreted as firms purchasing inputs, we obtain a model where firms might consider vertically integrating (or signing long-term contracts) to better assure themselves of supply at a lower than market price. Any firm with certain demand could supply its own needs at a price below the market price. In the markets under study, externalities can occur when supplying firms cannot distinguish among customers with differing demand uncertainties. Firms might vertically integrate to escape paying for the costs that someone else's uncertainty imposes on the market. (See the author, 1976, for an examination of vertical integration.)

This paper emphasizes that the behavior of markets characterized by price inflexibility, demand uncertainty, and production lags differs in important respects from that of traditional deterministic markets. Because of the prevalence of these "nontraditional" markets in the economy, their further study definitely seems warranted.

APPENDIX

A

Define the expected shortage M for one firm with customer capacity s as

$$M(s) = \sum_{s+1}^{\infty} (i - s)pr(i)$$

$$\text{or } M(s) = \sum_{s+1}^{\infty} (i - s)pr(i) + (\bar{s} - s)pr(i)$$

$$\text{or } M(s) \doteq -\sigma N^L(u) - u\sigma(1 - F(u))$$

$$\text{where } u = (s - \bar{s})/\sigma, \bar{s} = L/N,$$

$$\sigma = \sqrt{\frac{L}{N}},\ N^L(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} te^{-1/2t^2}dt$$

$\bar{s} = E(i)$, $pr(i)$ = binomial probability that i of the L customers come to one store. Notice that $\bar{s}$ and σ^2 are the mean and approximate variance of this binomial process. Hence $(i - \bar{s})/\sigma$ is approximately normally distributed with mean 0, and variance 1.[31] If all firms follow the same operating policy then

$$1 - \lambda(s) = 1 - N \cdot M/L \text{ or}$$

$$1 - \lambda(s) = 1 - M/\sigma^2 \text{ or}$$

$$1 - \lambda(s) = [\sigma^2 + \sigma N^L(u) + \sigma u[1 - F(u)]]/\sigma^2$$

or defining $I(s) = N^L(u) - uF(u)$, we have $1 - \lambda(s) = (\sigma I + s)/\sigma^2$.

B

PROOF of Theorem 2:

The Lagrangian for the government's maximization problem can be written as

$$\text{(A1)}\quad \mathcal{L}(p, T, s, \mu) = (1 - \lambda)V(p, Y + T) + \lambda U(0, Y + T) - \mu\left[\left((1 - \lambda)p\frac{L}{N} - cs\right) x(p, Y + T) - \frac{L}{N}T\right]$$

where μ is a negative Lagrange multiplier.[32]

[31] See Feller, ch 7.

[32] The notation was defined beneath (6). Recall that λ is not a Lagrange multiplier, but is the probability of disappointment which is a function of s given in (3).

The first-order conditions are:[33]

$$(A2)\quad (1-\lambda)V_1 = \mu\left[(1-\lambda)\frac{L}{N}x + \left[(1-\lambda)\frac{L}{N}p - sc\right]x_1\right]$$

$$(A3)\quad (1-\lambda)V_2 + \lambda U_2 = \mu\left[\left[(1-\lambda)\frac{L}{N}p - sc\right]x_2 - \frac{L}{N}\right]$$

$$(A4)\quad \frac{d(1-\lambda)}{ds}\frac{N}{L}[V-U] = \mu\left[\frac{d(1-\lambda)}{ds}p - c\right]x$$

$$(A5)\quad \left[(1-\lambda)\frac{L}{N}p - cs\right]x = \frac{L}{N}T$$

Substituting (A5) into (A2) and (A3), we obtain

$$(A6)\quad (1-\lambda)V_1 = \mu\left[(1-\lambda)\frac{L}{N}\cdot x + \frac{L}{N}\frac{T}{x}x_1\right]$$

and

$$(A7)\quad (1-\lambda)V_2 + \lambda U_2 = \mu\left[\frac{L}{N}\frac{T}{x}x_2 - \frac{L}{N}\right]$$

Since V is an indirect utility function, we have that $x = -V_1/V_2$. Using this relation, rewrite (A6) as

$$(1-\lambda)V_1 = \mu\left[(1-\lambda)\frac{L}{N} + \frac{L}{N}\frac{T}{x}\frac{x_1}{x}\right]\cdot\left(-\frac{V_1}{V_2}\right)$$

or

$$(1-\lambda)V_2 = (-\mu)\left[(1-\lambda)\frac{L}{N} + \frac{L}{N}\frac{T}{x}\cdot\frac{x_1}{x}\right]$$

or

$$(A8)\quad V_2 = (-\mu)\left[\frac{L}{N} + \frac{L}{N}\frac{T}{x}\cdot\frac{x_1}{x}\frac{1}{1-\lambda}\right]$$

From (A7) and the above assumption, it follows that

$$(A9)\quad \mu\left[\frac{L}{N}\frac{T}{x}x_2 - \frac{L}{N}\right] < V_2$$

Substituting the expression for V_2 from (A8) into (A9), we have that

$$\mu\left[\frac{L}{N}\frac{T}{x}x_2 - \frac{L}{N}\right] < (-\mu)\left[\frac{L}{N} + \frac{L}{N}\frac{T}{x}\frac{x_1}{x}\frac{1}{1-\lambda}\right]$$

or

$$(-1)(-\mu)\frac{L}{N}\frac{T}{x}x_2 - \mu\frac{L}{N} < -\mu\frac{L}{N} - \mu\frac{L}{N}\frac{T}{x}\frac{x_1}{x}\frac{1}{1-\lambda}$$

or since $-\mu > 0$,

$$(A10)\quad (-1)Tx_2(1-\lambda) < \frac{T}{x}x_1$$

$$\text{or}\quad T\left(-\frac{x_1}{xx_2}\right) < T(1-\lambda)$$

If $T > 0$, then $-x_1/xx_2 < 1-\lambda < 1$, while if $T < 0$, then $-x_1/xx_2 > 1-\lambda$. But from the Slutsky equation, we know that $x_1 + xx_2 \leq 0$ or $-x_1/xx_2 \geq 1$. Therefore if $T > 0$, we obtain a contradiction. Hence only $T < 0$ is possible in the optimal solution.[34,35]

[33]Subscripts denote partial differentiation

[34]I am implicitly assuming that the optimal solution is an interior solution and satisfies the first- and second-order conditions. Corner solutions ($T = -Y$ or $s = 0$) are assumed not to be optimal.

[35]If we assume that $x_2 > 0$ and $V_2(p, Y) = u_2(0, Y)$, then (A10) (with an equality sign) implies that the optimal $T = 0$. This result follows from the fact that if $x_2(1-\lambda)$ equaled $-x_1/x$, the Slutsky condition $x_1 + x \cdot x_2 \leq 0$ would be violated.

REFERENCES

A. **Alchian,** "Information Costs, Pricing, and Resource Unemployment," in Edmund Phelps, ed., *The Microeconomic Foundations of Macroeconomic Theory*, New York 1970.

G. **Brown, Jr. and M. Johnson,** "Public Utility Price and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119-28.

G. **Butters**, "Equilibrium Price Distributions," *Rev. Econ. Stud.*, Oct. 1977, *44*, 465-92.

D. **Carlton**, "Market Behavior under Uncertainty," unpublished doctoral dissertation, Mass. Instit. Technology 1975.

______, "Vertical Integration in Competitive Markets under Uncertainty," work. paper no. 174, econ. dept., Mass. Instit. Technology, Apr. 1976.

______, (1977a) "Pricing, Uncertainty, and Production Lags," *Amer Econ. Rev Proc.*, Feb. 1977, *67*, 244-49.

______, (1977b) "Peak Load Pricing with Stochastic Demand," *Amer Econ. Rev.*, Dec. 1977, *67*, 1006-10.

Alfred Chandler, Jr., *Strategy and Structure, Chapters in the History of American Industrial Enterprise*, Cambridge 1964.

A. De Vany and T. Saving, "Product Quality, Uncertainty and Regulation: The Trucking Industry," *Amer. Econ. Rev.*, Sept. 1977, *67*, 583-94.

William Feller, *An Introduction to Probability Theory*, Vol. I, New York 1968.

J. **Gould**, "Inventories and Stochastic Demand: Equilibrium Models of the Firm and Industry," *J Bus., Univ Chicago*, Jan 1978, *51*, 1-42.

J. **McCall**, "Probabilistic Microeconomics," *Bell J. Econ.*, Fall 1971, *2*, 403-33.

Edwin Mills, "Uncertainty and Price Theory," *Quart. J Econ.*, Feb. 1959, *73*, 116-30.

______, *Prices, Output, and Inventory Policy*, New York 1962.

S. **Rosen**, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34-55.

M. **Rothschild**, "Models of Market Organization with Imperfect Information: A Survey," *J. Polit. Econ.*, Nov. 1973, *81*, 1283-308.

______ **and J. Stiglitz**, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quart J. Econ.*, Nov. 1976, *90*, 629-50.

______, "Models of Market Organization with Imperfect Information: A Survey," *J Polit Econ.*, Nov. 1973, *81*, 1283-308.

M. **Visscher**, "Welfare Maximizing Price and Output with Stochastic Demand: Comment," *Amer Econ. Rev*, Mar 1973, *63*, 224-29.

E. **Zabel**, "A Dynamic Model of the Competitive Firm," *Int. Econ Rev.*, June 1967, *8*, 194-208

Market and Shadow Land Rents with Congestion

By RICHARD J. ARNOTT AND JAMES G. MACKINNON*

It is well known that when there are imperfections in the economy, shadow prices should be used instead of market prices in cost-benefit analysis. A major imperfection in urban economies is that residents do not pay the social cost of the congestion they create. As a result, market and shadow land rents may be different. This paper investigates the relationship between the two. The conventional view, expressed by Robert Solow, is that the shadow rent on land in residential use always exceeds the market rent. One major result of our paper is that this conclusion is shown to be incorrect. We also examine the shadow rent on land in transportation, and show that it may be negative.

This subject is of considerable interest for policymakers. Many urban public expenditures, such as the construction of roads and government buildings and the creation of public recreational land, involve the acquisition of land. If efficient decisions are to be made on such expenditures, it is clearly necessary for governments to know the shadow rent on land. Also, when there are divergences between shadow and market rents, governments may want to intervene in the market to alter private decisions, which are socially inefficient since they are based on market rents. Governments may also want to change the pricing of transportation so as to reduce or eliminate the divergence between the private and social costs of congestion, which is responsible for the divergence between market and shadow land rents. William Vickrey (1963) has proposed various schemes which would do this. In Section III of this paper the income-equivalent benefit of charging (almost) optimal congestion tolls is computed using a numerical simulation model. To the extent that this model is realistic, we can provide some indication of the expenditure that would be justified to implement a Vickrey-type scheme.

Previous work on the relationship between market and shadow land rents is not entirely satisfactory. Solow and Vickrey investigated analytically the effects of employing an incorrect planning rule to determine road width in a model with a very restrictive technology, where all land is in commercial use. Yoshitsugu Kanemoto (1975, 1976) undertook similar analyses for models in which land use is industrial. Richard Muth (1975) used a simulation model to investigate whether too much or too little land is allocated to streets. Solow constructed a simulation model which permitted him to investigate whether, on average, shadow land rents exceed market land rents on land in residential use, but not how shadow and market rents are related as a function of location. Kanemoto (1977) investigated the relationship between shadow and market rents on residential land for the case where road widths are optimal given that congestion is not priced. Since it is doubtful that roads are approximately of second best optimal width, his results are not generally applicable. All these papers are, at least implicitly, concerned with the relationship between market and shadow land rents, but this is, to our knowledge, the first paper to deal explicitly with that relationship as a function of location, in a residential location theory model which is reasonably realistic.

In Section I we investigate the analytical relationship between the market and shadow rent on residential land in the presence of unpriced congestion. We use a model similar to Solow's. The conventional argument concerning this relationship is shown to be

*Queen's University. We would like to thank Ronald Grieson, William Vickrey, an anonymous referee, and participants of the Queen's University Microeconomics Workshop for helpful comments. Remaining errors are our responsibility.

faulty. In Section II a numerically solvable residential location model and the technique used to solve it are described. In Section III this simulation model is used to investigate the relationship between the market and shadow rents on residential land and the shadow rent on land used for transportation, and to estimate the benefits from charging congestion tolls. In Section IV other factors affecting the relationship between market and shadow rents are discussed briefly.

I. The Relationship Between Shadow and Market Rents

Shadow rents can be computed in a number of different ways. In this paper, the following procedure is adopted. To ascertain the shadow rent on residential land at a particular location, add a small amount of residential land at that location, and solve for a new equilibrium using lump sum transfers to ensure that all city residents achieve their previous level of utility. The shadow rent is then the money saved by the government as a result of the land being added, divided by the amount of land that was added. Alternatively, a small amount of land could be subtracted, and the shadow rent would be the additional money spent by the government in order to keep residents at their previous utility level, divided by the amount of land that was subtracted. In the limit, of course, these two procedures give identical results.

A major conclusion is that when congestion is unpriced, the value to society from adding land to roads is, in general, *not* equal to the direct transport savings associated with the land addition. A (compensated) transport improvement or the (compensated) addition of residential land results in residents expanding lot size on average, and in an increase in traffic flow at every location. *This increased flow causes an increase in the excess burden associated with congestion not being priced.* The conventional argument ignores this, and by doing so incorrectly computes shadow rents. It comes to the erroneous conclusion that the shadow rent on residential land always exceeds the corresponding market rent (except at the boundary where they are equal). These points are now elaborated.

Consider a very simple residential location theory model similar to those dealt with by Solow and Kanemoto (1977). The city has a fixed population of households with identical tastes and incomes, who derive satisfaction from land and from other goods. All markets are competitive, so that residents have equal utility in equilibrium. Every day, all residents must commute to and from the central business district (CBD). That is the only travel which occurs. Also, land is homogeneous. Thus, locations are differentiated solely on the basis of accessibility to the CBD, which is measured by x. A resident may consume land at only one location. The lot size of the resident at x is denoted by $T(x)$, market and shadow land rents by $r(x)$ and $s(x)$, respectively, nonland consumption of the resident at x by $C(x)$, and the population between x and $x + dx$ by $n(x)dx$. The boundary of the city is at x^b, and the land rent at the boundary, $r(x^b)$, is equal to $\bar{r}$, the exogenous rent on land in agricultural use.

There is only one transport artery (or, equivalently, many identical ones), so that the number of travellers on it at rush hour, $Q(x)$, equals the number of residents who live between x and the boundary of the city; i.e.,

$$Q(x) = \int_x^{x^b} n(x')dx' \tag{1}$$

Transport costs between x and $x + dx$ are $g[w(x), Q(x)]dx$, where $w(x)$ is the width of the road. The dependence of transport costs on w and Q reflects flow congestion in transportation. As the number of travellers on the road increases, so do transport costs ($g_Q > 0$), and as the width of the road increases, transport costs decline ($g_w < 0$).

Now consider what happens when an amount of residential land $T(x^*)$ is added to the residential land available at location x^*. This is exactly the amount of land occupied by each resident at x^*. *If it is assumed that lot sizes remain fixed*, the net

effect of the addition must be the movement of one resident from the boundary of the city to x^*. The shadow rent of the added lot, $s(x^*)T(x^*)$, is the income that the government can extract and still leave everyone with the same allocations they had before (or, in the case of the mover, with the same allocation as other residents at x^*). This is the sum of three components: the reduction in aggregate transport costs, the mover's reduction in expenditure on other goods (which is typically negative), and the income derived from the vacated lot at the edge of the city (which can now be used for agriculture).

The mover's transport costs are reduced by

$$(2) \qquad \int_{x^*}^{x^b} g\,dx$$

Since he no longer travels between x^b and x^*, the number of travellers on the road between x^b and x^* decreases by one, which reduces travel costs of residents between x^* and x^b. The reduction in transport costs between x and $x + dx$, where x lies between x^* and x^b, is

$$(3) \qquad Q(x)g_Q(x)dx$$

The transport costs of those living between 0 and x^* remain the same. Thus the total reduction in transport costs is

$$(4) \qquad \int_{x^*}^{x^b} g\,dx + \int_{x^*}^{x^b} Qg_Q dx$$

The income derived from the vacated lot is $r(x^b)\,T(x^b)$. The reduction in the mover's expenditure on other goods is

$$(5) \quad C(x^b) - C(x^*) = r(x^*)T(x^*) - r(x^b)T(x^b) - \int_{x^*}^{x^b} g\,dx$$

from the mover's budget constraints at x^* and x^b. Adding up these three terms yields

$$(6) \quad s(x^*)\,T(x^*) = \int_{x^*}^{x^b} Qg_Q dx + r(x^*)\,T(x^*)$$

Rearrangement of (6) then yields

$$(7) \quad (s(x^*) - r(x^*))\,T(x^*) = \int_{x^*}^{x^b} Qg_Q dx$$

Expression (7) indicates that the shadow rent on the added lot exceeds the market rent by an amount equal to the value of transport savings to the nonmovers resulting from the decrease in congestion between x^* and x^b. Since these savings must be positive for all x^* less than x^b, $s(x)$ must exceed $r(x)$ at every location, except at the boundary where they are equal. The slope of the market rent gradient reflects a resident's reduction in transport costs resulting from his moving from $x + dx$ to x. The slope of the shadow rent gradient, however, reflects the sum of the resident's reduction in transport costs and the reduction in nonmovers' transport costs resulting from the move. Since, with the specified assumptions, congestion is always reduced by such a move, the shadow rent gradient is always steeper than the market rent gradient at a particular location. Thus, the shadow rent exceeds the market rent by an increasingly large amount as distance to the city center is reduced.

The foregoing argument is correct, given the assumption that lot sizes remain fixed. However, the addition of land increases the supply of land in residential use. As a result, rents fall on average, causing residents to substitute land for other goods in consumption. Average lot size increases, and more people travel on the road at all locations, relative to the situation after the addition of land but before lot sizes adjust. Consequently, the amount of unpriced congestion at all locations increases, and also the excess burden associated with the unpriced congestion. By assuming that lot size is fixed, the conventional method of calculating the shadow rent on residential land yields a measure of the shadow rent that is consistently biased upwards by an amount equaling the increase in excess burden.

A similar argument can be made for the shadow rent on land used in transportation. If one assumes that lot sizes do not adjust in response to the addition of land to the road at some location x^*, the shadow rent on land in road use there is computed as

$$(8) \qquad -Q(x^*)g_w(x^*)$$

which is the number of road users at that lo-

cation, $Q(x^*)$, times the travel savings to each when the road is widened by one unit for a distance of one unit, $-g_w(x^*)$, or simply the direct transport savings associated with the addition of the land. But since it ignores lot size adjustment, which increases the excess burden associated with unpriced transportation congestion, this measure also consistently overstates the true shadow rent.[1]

II. A Simulation Model

The conventional argument suggests that the shadow rent on land in residential use always exceeds the market rent. In the preceding section it was argued that the shadow rent, correctly measured, is less than the shadow rent measured according to the conventional argument. The obvious question that arises is, can the shadow rent on residential land be less than the market rent? To answer this question, a simulation model which requires numerical solution is constructed. The model is described in this section, and the results of several simulations are presented in Section III.

This model is similar to Solow's, which was also solved numerically, but is substantially more complicated and realistic It incorporates three of his four suggested extensions: ". . (1) the explicit inclusion of housing in addition to land as a residential cost; (2) the allowance for both time costs and out-of-pocket costs of commuting; (3) the use of a congestion-cost function that rises more than proportionally with traffic density . . ." (p. 617). Solow's fourth suggestion, that two or more income classes be dealt with, was incorporated in one simulation run, but since this paper focuses on issues of efficiency rather than of distribution, only the one-group model is described here.

The city has a population of 1 million households, each of which has an annual after-tax income of $13,000, and an indirect utility function

$$U = Y P_h^{-.3} \tau \tag{9}$$

where Y is income net of money transport costs (inclusive of toll charges, where applicable), P_h is the rental price of housing, and τ is defined by

$$\tau = 1 - .125\, Time \tag{10}$$

where *Time* is the amount of time, in hours, it takes to get to the *CBD*. Thus if Y and P_h did not depend on the household's location, which of course they do, utility would decline linearly with the time it takes to get to the *CBD*, reaching zero at a travel time of eight hours. Equation (9) implies that the household spends 30 percent of its income net of transport costs on housing, and 70 percent on other goods. The functional form (10) is justified in the authors' article (1977b). What matters for the purposes of this paper is that utility declines, *ceteris paribus*, as the time spent commuting increases.

Housing is treated as a nondurable good, which is produced in a perfectly competitive market at each location. The rent on housing is determined by a constant elasticity of substitution (*CES*) cost function,

$$P_h = (.1 r^3 + .9 P_S^3)^{1/3} \tag{11}$$

where r is the rent on land, which varies across locations, and P_S is the rent on a unit of structure, which is assumed to be $1,300. The above cost function implies that the elasticity of substitution in the production of housing is .7, a figure in line with empirical work by Muth (1971) and Roger Koenker.

For purposes of numerical solution, the city is divided into a number of concentric rings each half a mile wide, around a *CBD* with a radius of one mile. One member of

[1] Cost-benefit analysts are concerned with values rather than rents Because of the durability of housing, it may take a long time for the city to adjust to a lot addition or the widening of a road In a stationary economy, value may usefully be viewed as a weighted average of the capitalized value of long-run rents (we compute long-run rents) and the capitalized value of short-run rents (By implicitly assuming fixed lot sizes, Solow computes short-run rents) The weight depends on the speed of adjustment of the economy The various qualitative propositions we develop concerning rent relationships apply also to value relationships.

every household which lives in ring i is assumed to commute from the midpoint of the ring, making 250 round trips or 500 one-way trips per year. Since money transport costs are assumed to be 15¢ per mile, regardless of congestion, the net income of a household which lives in ring i is

$$(12)\quad Y = 13{,}000 - (500)(.15)\cdot(1.0 - .25 + .5i) = 12{,}943.75 - 37.5i$$

Dividing the city into discrete rings inevitably introduces some inaccuracy, but since the maximum difference between where a household lives and where it is assumed to live is only one quarter of a mile, this is quite small.

The treatment of congestion in existing urban models is not very satisfactory. Only flow congestion is considered. Congestion associated with intersections and with entry to and exit from the traffic flow, and queuing phenomena are ignored. The latter may be important, and may have striking implications for the relationship between shadow and market land rents. Dan Usher has developed a location theory model in which the only form of congestion is queuing congestion, and obtained the result that the market rent on land *always* exceeds the shadow rent. Consider a very simple version of his model, in which the city is a long narrow parking lot. At one end of the parking lot is a gate, through which traffic flow is limited. Every morning residents get in their cars and wait their turn to go through the gate into the *CBD*. Waiting is the only cost of travel. Clearly, residents will be willing to pay more for locations nearer the gate, because the length of their wait depends on their position in the queue. Thus market rents decline monotonically with distance from the gate. Now suppose that a new parking place is created. Its shadow rent is clearly just the opportunity rent on land in other uses, because moving someone from another space into the new one does not change aggregate congestion at all; it simply frees up a parking space elsewhere. The market rent on the last space in the queue is also the opportunity rent on land in other uses, which is equal to the shadow rent on every space in the queue. But the market rent on all spaces except the last exceeds the market rent on the last space. Thus in Usher's model, market rents always exceed shadow rents, except at the boundary where they are equal.

The conventional modelling of congestion in urban models not only ignores forms of congestion other than flow congestion; its treatment of flow congestion is also unsatisfactory. The assumed specifications of flow congestion imply that there is no maximum feasible flow; flow can always be increased at some cost in time. But traffic engineering studies (see Institute of Traffic Engineers, pp. 271–76) suggest that there is a maximum flow, and that if traffic tries to exceed it, flow is actually reduced and queuing must occur. To model this phenomenon realistically would be very difficult, since the length of the queues must vary with the time of day. Moreover, when a person enters the queue depends not only on where he lives but also on when he begins the journey to or from work, which should be determined endogenously. Because of these difficulties, we have chosen with some reluctance to follow the conventional modelling of congestion. It should be emphasized that the results obtained in this paper are contingent on the treatment of congestion.

The specification of the flow-congestion function is now considered. Most authors do not distinguish between the time and money costs of commuting. Observation suggests, however, that congestion has a large effect on time costs, but only a small effect on money costs. It is assumed here that money costs are unaffected by congestion, but that time costs are related to flow per unit width of road. Specifically,

$$(13)\qquad t(x) = t[Q(x)/w(x)]$$

where $t(x)$ is the time required to travel a mile at x, $Q(x)$ is the number of households living beyond x, and $w(x)$ is the width of the road in feet times forty. Width is measured in this way so that $Q(x)/w(x)$ equals unity

with "normal" traffic flow. The factor of 40 enters since the rush hour is assumed to be one hour long, and since normal traffic flow is defined to be forty cars per hour per foot of road width.

A number of additional stylized facts were employed in choosing the specific functional form of (13). Before introducing them, it is necessary to develop some terminology. Define the private congestion cost at x, $PC(x)$, to be the increase in the time required for an individual to cover a mile at x due to the presence of others on the road; i.e., $PC(x) = t[Q(x)/w(x)] - t[0/w(x)]$. Total travel time on a mile of road at x is $t(x)Q(x)$. When one more traveller is added, the increase in total travel time is

$$(14) \qquad t(x) + t_Q(x)\,Q(x)$$

The first term in (14) is the private cost to the traveller; the second term, the congestion externality (in time units) imposed by the marginal traveller on other persons on the road. This latter term is defined to be the marginal congestion externality at x, $MCE(x)$.

The additional stylized facts that we considered were that:

(i) free flow speed be reasonable;

(ii) the elasticity of private congestion with respect to flow exceed unity $(\partial PC/\partial Q)(Q/PC) > 1$); and

(iii) the ratio of the marginal congestion externality to private congestion should be an increasing function of Q, where[2]

$$\partial\left(\frac{MCE}{PC}\right)\Big/\partial Q > 1$$

The flow congestion function employed in this paper is

$$(15) \quad t(x) = t_0 \exp\left[\alpha(Q(x)/w(x))^{\beta}\right] \qquad \beta \geq 1, \alpha > 0, t_0 > 0$$

It is easy to show (see the authors, 1976) that this functional form satisfies the above stylized facts; however, the flow congestion function used most commonly in the literature,[3]

$$(16) \quad t(x) = t_0 + k(Q(x)/w(x))^{a}, \qquad a > 0, k > 0, t_0 > 0$$

does not. Thus (15) is the better functional form.

In the simulation runs, t_0 was set at 1/35, which implies a free flow velocity of 35 miles per hour; the other two parameters of (15) were chosen to result in a reasonable speed gradient for rush hour traffic, and were varied over the simulation runs. The amount of land devoted to streets was also varied over the simulation runs, while the amount of land devoted to housing was set equal to half the potentially available land in each ring beyond the *CBD*. This reflects the fact that much land in real cities is used for purposes other than housing and roads.

The model was solved using a variant of the technique employed by the authors (1977a,b), which makes use of a simplicial search algorithm called the Vector Sandwich Method (see MacKinnon) which is similar to the algorithms developed by Herbert Scarf. At each iteration, the algorithm "tries out" a rental price of housing at the center of the city. From this can be calculated the utility level achievable by a household living at the city center, which must be the utility level achieved everywhere else as well. Flow through the first ring is always 1 million people, so that time cost for households in the first ring can be calculated, and that, along with the utility level they must achieve, implies the rents on housing and land they must face, which in turn determines how many people can fit in the first ring. Flow through the second

[2]Solow argued for the use of (but did not himself employ) "a congestion-cost function that rises more than proportionally with traffic density . . ." (p. 617). Since Solow took $Q(x)$ as a proxy for density, rather than flow, we took his statement to mean that $(\partial PC/\partial Q)(Q/PC) > 1$. Thus (ii) is Solow's stylized fact. Vickrey (1969) implies that the ratio of *MCE* to *PC* should be an increasing function of Q. Thus (iii) is Vickrey's stylized fact.

[3]Actually, the flow congestion function in the literature is of this form, but treats money expenditure rather than time expenditure.

ring can now be determined, hence its prices and population, and so on through the rings. The algorithm then searches for a rental price of housing at the center such that, when all residents are housed, the rent on urban land at the edge of the city is equal to the predetermined rent on agricultural land of $750 an acre.

In order to compute the shadow rent on residential land as a function of location, the model is first solved as above, and the level of utility achieved in this base solution stored. The model is then solved a number of times, with 100 acres of residential land added to each even-numbered ring in turn (odd-numbered rings were not included to save computer time).[4] In these simulations, utility is prespecified to equal the level achieved initially. The algorithm varies lump sum transfer payments from the government so that this is achieved, as well as varying the rent on housing at the center so that, when all residents are housed, the land rent at the edge of the city equals the agricultural land rent. The shadow rent on residential land in ring i is then calculated as

$$(17) \quad (-TP + \Delta DLR + 75{,}000)/100$$

where TP is the total transfer payments by the government to the households, ΔDLR is the change in differential land rents between the base solution and the compensated solution with added land, and 75,000 is the value of the added land in agricultural use (100 acres at $750 per acre) Note that the government is treated as owning all the land in the city. This is equivalent to assuming absentee landlords, who are also compensated in the new equilibrium (so that the land rents they receive are unchanged).

The shadow rent on land in transportation is calculated the same way as the shadow rent on land in residential use, except that only 50 acres are added to the land used for roads in every even-numbered ring in turn. Since there were generally around 22 occupied rings, calculating the shadow rents on land in transportation and in residential use for all even-numbered rings, required that the simulation model be solved around twenty-three times. This required about four minutes of processing time on a Burroughs B6700.

III. Simulation Results

In the first subsection, results are presented for the base case city; in the second, the effects of some alternative specifications of the road width and congestion functions are considered; and in the third, the efficiency gain from imposing a congestion toll is evaluated.

A. *The Base Case City*

In the base case city, the parameters of the congestion function (15) are $\alpha = .6$ and $\beta = 1.0$ (in all runs $t_0 = 1/35$). The characteristics of this function were discussed in the last section. At the normal flow where $Q/w = 1$, private congestion is 1.41 minutes time loss per mile, and the marginal congestion externality is 1.87 minutes time loss per mile Thus MCE/PC is 1.33 for $Q/w = 1$; for $Q/w = 0$ it is 1.00, and for $Q/w = 2$ it is 1 71.

The road width function, which gives forty times the width of the road in feet, as a function of location, is

$$(18) \quad w(x) = 211{,}200(.1x + .5)\pi$$

Thus the width of the road increases with x. Since the city is circular, the proportion of land used for the road is $(.05 + .25/x)$, which decreases with x. This function seems roughly realistic. 30 percent of the land is used for roads at the edge of the CBD, compared with 10 percent five miles from the center of the city.

Some of the characteristics of the simulated city are presented in Table 1. The

[4] The choice to add 100 acres rather than say, to subtract 50, is purely arbitrary. One hundred acres is small relative to the land available in any ring (about 6 percent of the land available in ring 2, and about 8 percent of the land available in ring 24), but not so small that round-off errors become important in the evaluation of (17) We experimented with other values, and found the calculated shadow rents quite insensitive to the choice, and certainly more than accurate enough for the uses to which they will be put

TABLE 1—CHARACTERISTICS OF THE BASE CASE CITY[a]

	First Ring	Middle Ring	Boundary Ring
Distance from the city center in miles	1.25	6.25	11.75
	(1.25)	(6.25)	(11.25)
Land rent per acre	19540	4456	768
	(34048)	(4257)	(771)
Structural density	1.377	.489	.143
	(2.032)	(.474)	(.143)
Housing price, 1,000 square feet equivalent	1928	1504	1238
	(2171)	(1495)	(1238)
Housing quantity, 1,000 square feet equivalent	2.008	2.500	2.937
	(1.848)	(2.506)	(2.923)
Land share in housing	.2003	.1385	.0867
	(.2284)	(.1369)	(.0867)
Speed (mph)	8.24	23.21	34.46
	(8.24)	(24.47)	(34.46)
Cumulative travel time in hours per trip	.0304	.3809	.5689
	(.0304)	(.3663)	(.5335)
Proportion of land used for roads	.2500	.0900	.0713
	(.2500)	(.0900)	(.0713)
Cumulative toll ($/year)	0	0	0
	(74.35)	(590.53)	(636.65)

[a] Numbers in parentheses refer to the same city with a congestion toll imposed.

numbers in parentheses refer to the same city with a congestion toll imposed (see Section IIIc). The base case city is in most respects quite similar to actual cities with 1 million households (for example, Toronto). The radius of the city is 11.75 miles. Structural density (square feet of floor area per square foot of residential land) varies from 1.377 at the center to 0.143 at the boundary of the city, a factor of about 10. Land rent varies by a factor of about 25, while the rent on housing varies by a factor of about 1.5. All these gradients are convex, as expected. Traffic speed increases from 8.24 miles per hour in the innermost ring to 34.46 miles per hour in the boundary ring, while cumulative travel time reaches about 35 minutes per trip at the boundary.

Figure 1 shows the relationship between the shadow rent on land in transportation computed correctly (curve I), the shadow rent on residential land computed correctly (curve II), and the market rent on residential land (curve III). Curves I and II differ because the road width is not second best optimal; comparison of those two curves suggests that road width is less than optimal up to about 6 miles from the edge of the CBD, beyond which it is greater than optimal.[5] Certainly the road is too wide at the edge of the city, since traffic flow goes to zero there, as does the shadow rent on land in transportation.

More interesting is the comparison between curves II and III. Curve II lies above curve III from the center of the city to ring 8, but below curve III from ring 9 to the boundary of the city; that is, *the market rent on residential land is less than the corresponding shadow rent in the inner section of the residential area, but exceeds the shadow rent in the outer section.* An intuitive explanation of this is the following. Compare the effects of adding a lot at x_1, which is near the center of the city where there is considerable transportation congestion, equal in size to existing lots at x_1, to the effects of adding

[5] This statement is only true in the sense that widening the road at a location where the shadow rent on land in road use exceeds the shadow rent on land in residential use will result in a Pareto improvement when no other changes are made. If the road width is altered at other locations, the shadow rents on land in both road and residential use will change at this location, so that, when road width at other locations has been adjusted optimally, it may actually be desirable to narrow the road at this location.

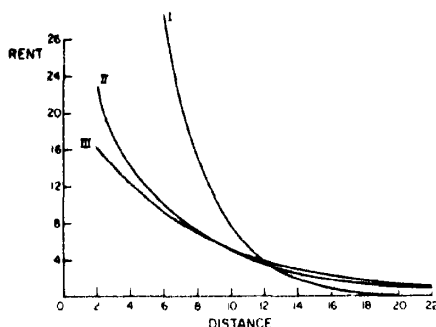
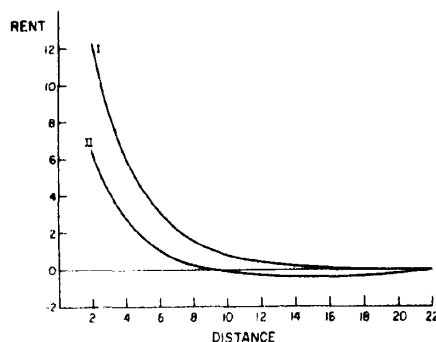



FIGURE 1

FIGURE 2

a lot at x_2, which is near the boundary of the city where congestion is negligible, equal in size to existing lots at x_2. Note that this implies that the amount of land added at x_2 must be considerably greater than the amount added at x_1. We have argued that the difference between the shadow rent and market rent on a residential lot can be separated into two components: first, the reduction in transport costs, induced by the lot addition, beyond the location where the lot was added, lot sizes fixed, which by itself would cause the shadow rent to exceed the market rent; and second, the effects resulting from general equilibrium lot size adjustment, which by themselves would cause the shadow rent to be less than the market rent. From (7), the first effect is larger the nearer is the lot to the city center because more people's travel costs are reduced. The magnitude of the second effect depends on how much people will expand their lots as a result of the addition of the lot; this can be expected to be more or less the same whether the lot is added near the center of the city or near the boundary. On balance then, one would expect the shadow rent on a lot minus its market rent to fall the further is the lot away from the city center. Furthermore, since the first effect becomes negligible near the boundary of the city, while the second does not, one would expect the market rent on a lot near the boundary to exceed the corresponding shadow rent.

Figure 2 shows the difference between the shadow and market rent on residential land as a function of location. Curve I shows this relationship when shadow rents are computed assuming lot size fixed; curve II, when shadow rents are computed correctly. The figure demonstrates the conventional result that the shadow rent on residential land computed assuming lot size fixed always exceeds the market rent. Also, in this simulated city, the shadow rent computed on the assumption of fixed lot size everywhere overstates the true shadow rent on residential land. The same is true of the shadow rent on land in transportation (not shown).

B. *Other Cities*

A number of other simulations were performed using different parameter values. In most cases, they yielded results qualitatively similar to those of the base case. One somewhat bizarre city did yield a new result, however.

This city differed from the base case in two ways. First, the parameters α and β of the congestion function were 0.2 and 3.0, respectively. With these parameters, MCE/PC is 3.00 for $Q/w = 0$, 3.33 for $Q/w = 1$, and 6.02 for $Q/w = 2$. The effect of these parameter changes is to make speed less sensitive to increases in flow for flows less than normal flow, and more sensitive to such increases for flows greater than normal flow. Second, the road width function was

$$(19) \quad w(x) = \begin{cases} 211{,}200\,(.603 - .03768x)\pi & \text{for } x \leq 5.75 \\ 211{,}200\,(.7537)\pi & \text{for } x > 5.75 \end{cases}$$

This road narrows to a distance 5.75 miles from the city center, and then suddenly widens and remains of constant width.

With these congestion and road width functions, this city is extremely congested in the inner portion of the residential area. Speed in the first residential ring is only 0.64 miles per hour. Congestion gradually decreases as one moves away from the center, until 5.75 miles away speed is 16.71 miles per hour. Then traffic speed jumps up as a result of the sudden widening of the road, so that at 6.25 miles from the center it is 34.70 miles per hour, nearly free flow velocity. Travel time to the boundary is 108 minutes. In the innermost ring, MCE/PC is 12.2: for every minute a traveller loses due to congestion, he causes others to lose 12.2 minutes.

With such a bizarre city, one should expect unusual and extreme results, and this is indeed the case. Figure 3 is comparable to Figure 1. The most remarkable feature of this city is the gradient of the shadow rent

FIGURE 3

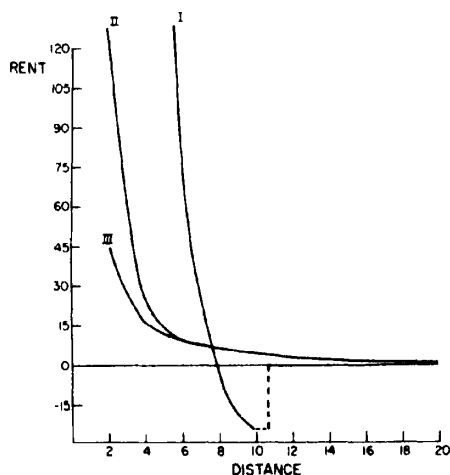
on land in transportation. It is roughly \$3 million per acre 0.75 miles from the inner residential boundary, and falls sharply until it is −\$24,389 per acre just before the road widens. Thus, *the shadow rent on land in transportation may be negative.* This can come about because the increased congestion elsewhere caused by widening the road at one location may exceed the direct savings.

At this point, it is useful to review what the simulation runs discussed in this subsection and the preceding one have shown. They have shown that:

1) the shadow rent on residential land correctly computed may be less than the shadow rent on residential land computed assuming lot size fixed (see Figure 2);

2) the market rent on residential land may exceed the shadow rent in quite realistic circumstances (see Figures 1 and 2);

3) the shadow rent on land in transportation correctly computed may be less than the shadow rent on land in transportation computed assuming lot size fixed (no figure, but implied by Figure 3); and

4) the shadow rent on land in transportation may be negative (see Figure 3).[6]

[6]Subsequent to the writing of this paper, Arnott analytically solved what was characterized in this paper as the Solow model (which is simpler than the simulation model in this paper). His results support the verbal arguments in the text of this paper, and suggest that most of the qualitative characteristics of the simulation runs presented here hold generally (contingent on the modelling of congestion). His results are summarized as follows: i) The measures of shadow rents computed ignoring lot size adjustment nearly always overstate the corresponding shadow rent computed correctly (nearly indicating a couple of cases where they are equal). ii) There is a critical location, x^*. Between x^* and the center of the city the shadow rent on residential land exceeds the corresponding market rent, while between x^* and the boundary the shadow rent on residential land is less than the corresponding market rent. iii) There is always a region near the boundary of the city where the shadow rent on land in road use is negative. iv) Between x^* and the center of the city, the difference between the shadow and market rent on residential land monotonically increases. It should be noted that while these results probably extend to the more detailed model treated in this paper, this has not been proved.

C. *Imposition of a Congestion Toll*

It would be difficult to impose an *optimal* congestion toll in the model of this paper. The problem is that congestion affects travel time, while any congestion toll must be in terms of money, and the shadow value of time varies with the location of the household. It is however possible to impose a congestion toll which is reasonably close to being optimal, certainly closer than tolls which realistically could be charged, by charging travellers the *MCE* they create times the shadow value of time of someone living at the center of the city (an arbitrary choice). If the congestion toll were really optimal, the shadow and market rents on residential land would be identical. With the toll that was imposed, the divergence between the two was always less than 10 percent, so that the toll is reasonably close to being optimal.

The toll was imposed on the base case city, and residents were compensated so as to keep them at their pretoll utility levels. Thus the efficiency gain is simply the change in government revenue. The characteristics of the base case city with the congestion toll are shown in parentheses in Table 1. Imposition of the toll increases the curvature of the rent and structural density gradients, and makes the city smaller and denser. Cumulative travel time to any point in the city is reduced as a result of the toll (but not by very much). The congestion toll per mile falls off very sharply away from the center of the city. A trip to the middle of the first ring costs 15 cents (or 60 cents per mile), while a trip to the boundary ring costs $1.27 (or 12.4 cents per mile).

The magnitude of the efficiency gain from the imposition of the toll is $8,801,761, or approximately $8.80 per household per year. Since the toll is not quite optimal, this is a lower bound. However, since excess burden tends to increase more than proportionately with the size of a distortion, it seems doubtful that the efficiency gain with an optimal congestion toll would be significantly larger than this.

For a model in which the average household is paying over $500 in tolls each year (see Table 1), $8.80 is a remarkably small figure for the efficiency gain from a congestion toll. The reason for this is that this model allows households very limited opportunities for escaping the toll. Every household must send someone to work in the *CBD* during the rush hour, so that congestion in the first ring is unchanged by the toll. People cannot avoid the toll by abandoning unnecessary trips, taking less congested roads, shifting trips to nonpeak hours, or taking jobs outside the *CBD*. All they can do is move their residences so that, on average, trips to work are shorter and cause less congestion. In view of this, we would caution strongly against attaching any practical significance to the figure of $8.80. Its smallness probably reflects the deficiencies of the treatment of congestion in residential location theory more than the deficiencies of congestion tolls. An accurate measure of the efficiency gains from congestion tolls could be computed by a technique similar to ours, but the model would have to be a great deal more complicated.

IV. Realistic Complications[7]

In a simple residential location theory model, Solow argued that in the long run the shadow rent on residential land exceeds the market rent. This paper has shown that this argument is incorrect, and has provided counterexamples. The actual relationship between market and shadow rents is considerably more complex than that implied by the conventional analysis. This conclusion would be reinforced if one were to introduce realistic complications.

One such complication is the presence of other distortions. In the models discussed above, unpriced transportation congestion was the only distortion in the urban economy. But other distortions, such as inefficient zoning and the property tax, are also important. The addition of land at some location may either increase or de-

[7]Ronald Grieson brought several of these points to our attention

crease the excess burden associated with these distortions, so that shadow rents may be either lower or higher than in the case analyzed in this paper.

There are many other features of the urban economy which one would want to capture in a model that was used for evaluating policy. These include the dynamic features caused by the durability of structures, migration into and out of the city, the complications of multiple workplaces determined endogenously, the very complex nature of congestion (which has been discussed in this paper), and the possible benefits associated with the open space afforded by urban roads.

V. Conclusions

This paper has investigated the relationship between shadow and market land rents. It was argued that the conventional wisdom, which asserts that shadow rents on residential land always exceed market rents, and that shadow rents on land in transportation are always positive, is incorrect because it does not measure shadow rents correctly. On the contrary, the simulation results demonstrated that the shadow rent on residential land may be less than the market rent, and that the shadow rent on land in transportation may actually be negative. The simulation model which was employed is more realistic than models which have been used by other authors, incorporating as it does the major extensions suggested by Solow. In particular, the form of the congestion function accords better with the available stylized facts than previous functional forms. However, as the small efficiency gain from imposing a congestion toll suggested, there are still many aspects of the model which are unrealistic. It is particularly necessary that further work be done on modelling congestion.

The results of the paper have a number of interesting policy implications. The correct calculation of the shadow rent on land in transportation and in residential use is evidently very difficult. However, this does not argue for the use, in cost-benefit analysis, of shadow rents computed assuming that lot size is unaffected by the addition of land, since these may be grossly incorrect; rather, the implication is that expenditure is justified in the development of urban simulation models to permit calculation of the true shadow rent with reasonable accuracy. An argument based on the conventional and incorrect calculation of shadow rents is that if planners allocate land to roads up to the point where the shadow rent on land in road use equals the market rent on residential land, then since the shadow rent on residential land exceeds the market rent, too much land is allocated to roads. However since the shadow rent on residential land correctly computed does not necessarily exceed the market rent, this argument is false.

REFERENCES

R. J. Arnott, "Unpriced Transportation Congestion," unpublished paper, Queen's Univ. 1977.

______ **and J. G. MacKinnon,** "Market and Shadow Land Rents with Congestion," disc. paper no. 250, Instit. Econ. Res., Queen's Univ. 1976.

______ **and** ______, (1977a) "The Effects of the Property Tax: A General Equilibrium Simulation," *J. Urban Econ.*, Oct. 1977, *4*, 389–407.

______ **and** ______, (1977b) "The Effects of Urban Transportation Changes: A General Equilibrium Simulation," *J. Publ. Econ.*, Aug. 1977, *8*, 19–36.

A. K. Dixit, "The Optimum Factory Town," *Bell J. Econ.*, Autumn 1973, *4*, 637–51.

Y. Kanemoto, "Congestion and Cost-benefit Analysis in Cities," *J. Urban Econ.*, July 1975, *2*, 246–64.

______, "Optimum, Market and Second-best Land Use Patterns in a von Thunen City with Congestion," *Reg. Sci. Urban Econ.*, Feb. 1976, *6*, 23–32.

______, "Cost-benefit Analysis and the Second-best Land Use for Transportation," *J. Urban Econ.*, Oct. 1977, *4*, 483–503.

R. Koenker, "An Empirical Note on the

Elasticity of Substitution between Land and Capital in a Monocentric Housing Market," *J. Reg. Sci.*, Aug. 1972, *12*, 299-306.

J. G. MacKinnon, "An Algorithm for the Generalized Transportation Problem," *Reg. Sci. Urban Econ.*, Nov. 1975, *5*, 445-64.

R. F. Muth, "The Derived Demand for Urban Residential Land," *Urban Stud.*, Oct. 1971, *8*, 243-54.

———, "Numerical Solution of Urban Residential Land-use Models," *J. Urban Econ.*, Oct. 1975, *2*, 307-32.

Herbert E. Scarf, with the collaboration of Terje Hansen, *The Computation of Economic Equilibria*, New Haven 1973.

R. M. Solow, "Congestion Cost and the Use of Land for Streets," *Bell J. Econ.*, Autumn 1973, *4*, 602-18.

——— **and W. S. Vickrey,** "Land Use in a Long Narrow City," *J. Econ. Theory*, Dec. 1971, *3*, 430-47.

D. Usher, "A Theorem about Urban Land Values and the Social Cost of Greenbelts," unpublished paper, Queen's Univ. 1976.

W. S. Vickrey, "Pricing in Urban and Suburban Transport," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 452-65.

———, "Congestion Theory and Transport Investment," *Amer. Econ. Rev. Proc.*, May 1969, *59*, 251-60.

Institute of Traffic Engineers, *Transportation and Traffic Engineering Handbook*, Englewood Cliffs 1976.

Devaluation, Wealth Effects, and Relative Prices

By HARVEY LAPAN AND WALTER ENDERS*

The emergence of the portfolio balance approach[1] has led to a reformulation of the causes of balance-of-trade and payments disequilibria. According to this approach, balance-of-trade deficits and surpluses reflect discrepancies between desired and actual wealth holdings; while balance-of-payments deficits and surpluses reflect discrepancies between desired and actual money holdings. Thus, balance-of-trade and payments disequilibria are viewed as representing disequilibria within the asset markets. Using this framework several authors[2] have examined the self-correcting nature of disequilibria within the balance-of-payments accounts and the ability of a devaluation to reduce the magnitude of a disequilibrium.

Two recent examples of this approach that have appeared in this *Review* are a paper by Rudiger Dornbusch and a paper by Jacob Frenkel and Carlos Rodriguez. The Dornbusch paper analyzes a two-country world in which each country issues a fiat money, while the Frenkel and Rodriguez paper develops a small country, two-asset model.[3] In both papers, a devaluation is successful because it reduces real wealth in the devaluing nation and increases real wealth in the appreciating nation. In terms of the absorption approach, the devaluation reduces absorption via the cash balance effect.

While both of these papers are important contributions to the examination of the impact of a devaluation on trade balances, they leave some important questions unanswered. Neither paper is concerned with the efficacy of a devaluation when residents of a country hold assets denominated in terms of the foreign unit of account. In assuming that individuals only hold domestic currency denominated assets, each paper demonstrates that a devaluation will be successful if it acts to decrease real wealth in the devaluing nation. However, when residents of a country hold assets denominated in the foreign currency, an exchange rate change will impose capital losses on residents of the revaluing nation while residents of the devaluing nation will experience capital gains. Since a successful devaluation must reduce (increase) wealth in the devaluing (revaluing) nation, the efficacy of a devaluation is directly related to the extent to which domestics hold foreign currency denominated assets.

Another problem not addressed in these papers is how changes in the terms of trade affect the balance of trade. The prevailing view (see articles by S. C. Tsiang, Arnold Harberger, and Svend Laursen and Lloyd Metzler) is that, if a devaluation causes a nation's terms of trade to deteriorate, the efficacy of a devaluation is reduced. A reduction in the terms of trade leads to a decrease in the marginal propensity to save as individuals attempt to maintain their real standard of living. As individuals cannot maintain this standard of living forever, it is still necessary to clarify the impact of terms-of-trade changes when balance-of-trade disequilibrium is viewed as representing disequilibrium in the asset markets. An emerging view is that relative price changes have little significance in determining the

*Iowa State University and Institute for International Economic Studies, and Iowa State University, respectively. We would like to thank the managing editor for his helpful comments and suggestions.

[1]See Ronald McKinnon or Harry Johnson for the seminal articles on the portfolio or monetary approach.

[2]See Bijan B. Aghevli and George Borts, Enders or Donald Mathieson for an analysis of the self-correcting nature of balance-of-payments disequilibria.

[3]One of the assets in the Frenkel and Rodriguez paper is physical capital which is immobile. Claims on physical capital are, however, perfectly mobile across national boundaries such that the domestic and foreign interest rate must be equal. Further, their paper is restricted to the "small country" case, so that domestic prices rise by the amount of the devaluation.

size of the balance of trade. For example, Frenkel and Harry Johnson state:

> The accumulation or decumulation of assets depends on the aggregate relationship between domestic expenditure and income and does *not* depend on the composition of expenditure between exportables and importables, or between goods that, given the price structure, are classifiable into tradeable and nontradeable goods. Consequently, though relative price changes do influence the composition of expenditures, they play a secondary role in the monetary approach.... [p. 23]

Section I considers the case in which there are two traded goods and examines the roles of relative price changes and capital gains and losses in a devaluation. It is shown that relative price changes may be the only way a devaluation can improve the balance of trade when domestic holdings of foreign currency denominated assets are large. Section II considers the case of nontraded goods and, in contrast to the standard result, demonstrates that the efficacy of a devaluation is inversely related to the absolute values of the changes in the prices of nontraded goods. Our conclusions are presented in Section III. The Appendix of the paper considers the stability properties of the model.

I. Devaluation in a Two-Traded-Good World

A. *The Model*

The model we analyze is identical to that of Dornbusch, except we assume that there are two traded goods (X_1, X_2), and that residents of each country may desire to hold assets denominated in terms of the foreign unit of account (one asset is denominated in dollars, the other in pounds). Following Dornbusch, we assume that the *U.S.* (*U.K.*) demand for nominal wealth is a constant fraction of *U.S.* (*U.K.*) nominal income:

$$(1)\quad \overline{W} = kY = kP_1\tilde{Y};\qquad \overline{W}^* = k^*Y^* = k^*P_1^*\tilde{Y}^*$$

where P_i (P_i^*) = dollar (pound) price of good i
Y (Y^*) = dollar (pound) value of *U.S.* (*U.K.*) income
$\tilde{Y}$ $(\tilde{Y}^*)$ = real income in terms of good 1
$\overline{W}$ $(\overline{W}^*)$ = desired wealth holdings in dollars (pounds)
k (k^*) = desired ratio of wealth to income in the United States (United Kingdom)

Assuming the dollar price of pounds is e, commodity arbitrage implies

$$(2)\qquad P_i = eP_i^* \qquad i = 1,2$$

Under the assumption that each country produces both goods,

$$(3)\quad Y = P_1Q_1(\rho) + P_2Q_2(\rho);\qquad Y^* = P_1^*Q_1^*(\rho) + P_2^*Q_2^*(\rho)$$
$$\tilde{Y} = Q_1(\rho) + \rho Q_2(\rho);\qquad \tilde{Y}^* = Q_1^*(\rho) + \rho Q_2^*(\rho)$$

where $\rho \equiv P_2/P_1 = P_2^*/P_1^*$ = relative price of good 2
$Q_i(Q_i^*)$ = output of good i by the United States (United Kingdom)

and production in each country takes place along a concave production possibility frontier on which the output of each good depends only on relative prices.

Desired nominal expenditures (E, E^*) equal money income minus nominal desired saving (S, S^*):

$$(4)\qquad E = P_1\tilde{Y} - S;\qquad E^* = P_1^*\tilde{Y}^* - S^*$$

Following Dornbusch, we assume that desired saving is proportional to any discrepancy between desired and actual wealth:

$$(5)\quad S = \pi[kP_1\tilde{Y} - W];\qquad S^* = \pi^*[k^*P_1^*\tilde{Y}^* - W^*]$$

where π (π^*) = adjustment parameter
W (W^*) = actual *U.S.* (*U.K.*) dollar (pound) value of wealth

Assuming no net wealth creation by either government, the dollar value of the U.S. balance of trade is equal to U.S. wealth accumulation:[4]

$$(6) \qquad DW = B = -eDW^*$$

where B = U.S balance of trade
$Dx = dx/dt$

In short-run equilibrium total income must equal total expenditures, a condition which will be fulfilled if desired world saving equals zero. Further, actual wealth accumulation must equal desired saving in each country:

$$(7) \qquad S + eS^* = 0$$

$$(8) \qquad B = S = DW$$

As equation (7) does not preclude the possibility of an excess demand in one of the commodity markets and an equivalent excess supply in the other, equilibrium requires

$$(9) \quad Q_2(\rho) + Q_2^*(\rho) = D_2(\rho, E/P_1) + D_2^*(\rho, E^*/P_1^*)$$

where $D_2(D_2^*)$ is U.S. (U.K.) demand for good 2, which by standard assumptions is homogeneous of degree zero in prices and expenditures.

Given that total world saving equals zero (equation (7)) and that equilibrium prevails in the market for good 2, then total world demand for good 1 must necessarily equal the total world supply. Thus, given e, W, and W^*, equations (7) and (9) determine the equilibrium values of P_1 and ρ.

[4]We abstract from the interest rate effects due to the yield on assets. It is assumed that governments sterilize interest payments by imposing lump sum taxes (subsidies) equal in magnitude to the net interest receipts of domestics from abroad. Further, the amount of tax which any individual pays is not commensurate with that individual's asset holdings. Thus, income is equal to the value of domestic production and the change in wealth is equal to the balance of trade. Note that the sign of dB/de refers to the balance of trade and not the balance of payments. Footnote 10 discusses the effects of a devaluation on the balance of payments.

Substitute the two relations in (5) into equation (7) in order to solve for P_1 in terms of W, W^*, $\tilde{Y}$, $\tilde{Y}^*$. Substitute this expression into equation (8) to yield

$$(10) \quad B = \pi\pi^*[k\tilde{Y}(eW^*) - k^*\tilde{Y}^*W] \cdot [s\tilde{Y} + s^*\tilde{Y}^*]^{-1}$$

where $s = k\pi (s^* = k^*\pi^*)$ is the marginal propensity to save out of income

As can be seen from equation (10), a U.S. balance-of-trade deficit is a wealth phenomenon, that is, the United States will experience a trade deficit if the ratio of U.S. desired wealth to actual U.S. wealth is greater than the corresponding ratio for the United Kingdom.

B. *The Effects of a Devaluation*

From equation (10) it is readily seen that a devaluation of the dollar will alter the U.S balance of trade only insofar as it a) redistributes wealth among countries, or b) alters the ratio of U.S to U.K. real income (i.e., $\tilde{Y}/\tilde{Y}^*$). If, as Dornbusch assumes, no individual holds foreign denominated assets, it immediately follows that the devaluation redistributes wealth away from the devaluing country, thereby improving the balance of trade. Since a U.S devaluation increases (decreases) the outstanding dollar (pound) value of private wealth, it will lead to increases in dollar prices and decreases in pound prices. The extent to which prices actually rise will depend upon the percent of world wealth denominated in dollars, as well as the propensities to consume out of income and wealth. In any event, the rise (fall) of prices in terms of dollars (pounds) decreases real wealth in the United States and increases real cash balances in the United Kingdom. This redistribution of wealth acts to increase U.S. saving, decrease U.K saving, and improve the U.S. balance of trade.

To the extent that residents of a country hold assets denominated in terms of the foreign unit of account, the preceding anal-

ysis of a devaluation is faulty. The initial effect of a devaluation of the dollar will cause *U.S.* residents holding pound denominated assets to experience capital gains, while *U.K.* residents holding dollar denominated assets experience capital losses. Thus, a devaluation of the dollar may redistribute wealth towards the United States and away from the United Kingdom. The possibility of a perverse redistribution of wealth means that the proportion of foreign asset holdings in domestic portfolios will be an important determinate of the efficacy of a devaluation.

We assume that residents of each country hold some of their wealth in assets denominated in terms of the foreign currency. Denote the proportion of these holdings to total nominal wealth by m and m^*. Then the change in nominal wealth measured in local currency due to an exchange rate change is

$$(11)\quad \frac{dW}{de} = \frac{mW}{e} \qquad \frac{dW^*}{de} = -\frac{m^*W^*}{e}$$

Correspondingly, the change in nominal wealth measured in terms of foreign currency due to an exchange rate change is

$$(12)\quad \frac{d(W/e)}{de} = -\frac{(1-m)W}{e^2}$$

$$\frac{d(eW^*)}{de} = (1-m^*)W^*$$

From equations (11) and (12), it is seen that the percentage change in *U.S* wealth is m when measured in dollars and $-(1-m)$ when measured in pounds. For *U.K* residents, the percentage change in nominal wealth is $(1-m^*)$ when measured in dollars and $(-m^*)$ when measured in pounds. Thus *U.S.* wealth, measured in terms of either dollars or pounds, will fall relative to *U.K.* wealth only if $1-m-m^* > 0$. If $1-m-m^* < 0$, *U.S.* wealth will increase relative to *U.K.* wealth; and if $1-m-m^* = 0$, there will be no relative change in wealth. Recalling from equation (10) that $B \gtreqless 0$ as $k\tilde{Y}/k^*\tilde{Y}^* \gtreqless W/eW^*$, a devaluation of the dollar will redistribute wealth in the "wrong" direction if $1-m-m^* < 0$.[5]

Thus, if $1-m-m^* \leq 0$, a devaluation of the dollar must raise $\tilde{Y}$ relative to $\tilde{Y}^*$ if it is to be successful in increasing the balance of trade. Notice that $\tilde{Y}$ and $\tilde{Y}^*$ are both functions of only one variable (ρ), so that the devaluation can only be successful if it produces a change in relative prices and the nations have different supply or demand conditions.

Specifically, $d(\tilde{Y}/\tilde{Y}^*)/d\rho \gtreqless 0$ as $Q_2/\tilde{Y} \gtreqless Q_2^*/\tilde{Y}^*$. Thus, an increase (decrease) in the relative price of good 2 will act to improve (worsen) the *U.S.* balance of trade if the ratio of *U.S.* production of good 2 to *U.S.* real income is greater than the corresponding ratio for the United Kingdom. In the important special case in which it is possible to identify the exporter of a particular good as the nation which produces the largest amount of that good relative to its total production, a deterioration (improvement) in the terms of trade of the devaluing nation will act to worsen (improve) the balance of trade. In general, it should be clear that the impact of a relative price change on the ratio $\tilde{Y}/\tilde{Y}^*$ (and hence the balance of trade) depends upon the pattern of production (i.e., how much of each good a nation produces) and not upon the pattern of trade (i.e., the particular good a nation exports).[6]

The crucial point is that for a *U.S.* devaluation to be successful, it must increase *U S* saving One method to increase *U S*

[5]Individuals are assumed to have static expectations so that m and m^* can be treated as constants If individuals expect a devaluation of the dollar m will increase and m^* will fall, so that $(1-m-m^*)$ may change in either direction

[6]Note that this is not simply an index number problem which arises from our having defined real incomes in terms of good 1 As $d(\tilde{Y}/\tilde{Y})/dp \gtreqless 0$ as $Q_2/\tilde{Y} \gtreqless Q_2^*/\tilde{Y}^*$, $d[Y/P_2 + Y^*/P_2^*]/d(P_1/P_2) \gtreqless 0$ as $Q_2/\tilde{Y} \gtreqless Q_2^*/\tilde{Y}^*$ Thus, an increase in the relative price of good 2 will increase *U S* income relative to *U K* income when measured in terms of either good 1 or good 2 if the ratio of *U S* production of good 2 relative to its total production is greater than the corresponding ratio for the United Kingdom How the terms-of-trade effect alters the balance of trade then depends upon the pattern of production, and not upon the choice of numeraire

saving is through a relative transfer of wealth between nations. The other method is to increase the *U.S.* demand for wealth relative to that of the United Kingdom by increasing *U.S.* real income relative to *U.K.* real income. Relative price changes will act to increase *U.S.* income relative to *U.K.* income if the relative price change favors the good which the United States produces in a relatively greater proportion to its total income.

The relationships between the devaluation, the trade balance, and the terms of trade are formally obtained by differentiating equation (10) with respect to the exchange rate. Utilizing the relationships in (11) and (12), simplification yields

$$(13)\quad \frac{dB}{de} = \frac{mB}{e} + \frac{\pi^* W^*(s\tilde{Y})(1 - m - m^*)}{(s\tilde{Y} + s^*\tilde{Y}^*)} + a_0\left[\frac{Q_2}{\tilde{Y}} - \frac{Q_2^*}{\tilde{Y}^*}\right]\frac{dp}{de}$$

where a_0 is a positive number.[7] From equations (7) and (9),

$$(14)\quad sgn\left[\frac{dp}{de}\right] = sgn\,[1 - m - m^*][C_2^* - C_2]$$

where $C_2(C_2^*)$ is the *U.S.* (*U.K.*) marginal propensity to consume good 2. Equation (13) demonstrates that the terms of trade has an ambiguous effect on the trade balance (since $Q_2/\tilde{Y} - Q_2^*/\tilde{Y}^*$ may be positive or negative) while equation (14) demonstrates that the relative price of good 2 may increase or decrease. The latter follows as: if $1 - m - m^* > 0$ ($1 - m - m^* < 0$), wealth is redistributed towards the United Kingdom (United States). If wealth is transfered towards the United Kingdom, real *U.K.* expenditures will rise while real *U.S.* expenditures will fall. If the *U.K.* marginal propensity to consume good 2 is greater (less) than that of the United States, the relative price of good 2 will rise (fall). In the case in which the devaluation both redistributes wealth towards the country whose currency increases in value and in which countries tend to produce a large proportion of the good for which they have a high marginal propensity to consume, the change in relative prices will act to worsen the trade balance.

It can unambiguously be said that the greater the degree to which domestics hold assets denominated in terms of the foreign currency, the less effective is the devaluation.[8] To the extent that the holding of assets denominated in terms of the foreign unit of account is associated with the degree of capital mobility, the efficacy of a devaluation will be negatively related to the degree to which individuals view domestic and foreign assets as substitutes. If individuals view domestic and foreign assets as perfect substitutes, the expected values of m and m^* will be equal to 1/2 so that a devaluation will have no effect on the trade balance or on relative prices. In the case in which residents of a country only hold assets denominated in terms of their own unit of account, a devaluation always improves the trade balance: setting $m = m^* = 0$, we also find that in this special case, the devaluation always works. Lastly, if $1 - m - m^* < 0$, the devaluation will be counterproductive.

The crucial point to note is that when m and m^* are greater than zero, the initial gains (losses) of an exchange rate change act to offset the effects of price increases or decreases on wealth. This result implies that the impact of a devaluation cannot be divorced from the degree of asset substitutability. Those factors which induce

[7]The magnitude of a_0 depends upon supply and demand elasticities, but we are only interested in the direction of change.

[8]The condition that $1 - m - m^* > 0$ is not sufficient to guarantee that the devaluation improves the trade balance if the devaluing nation initially has a deficit. The larger the deficit, and the greater $m + m^*$, the less likely it is that the devaluation will succeed in improving the balance of trade, as measured in domestic currency units.

residents of a country to hold assets denominated in terms of a foreign unit of account—such as a large volume of trade or expectations of a devaluation by residents of the devaluing nation—act to work against using the exchange rate as a policy instrument.

The discussion above relates only to the impact effect of a devaluation, and has no bearing upon the stability of the system or the effects of a devaluation on the long-run values of the endogenous variables. In order to conserve space, we defer discussion of these problems to the nontraded goods case. In the nontraded goods case we demonstrate that a devaluation does not alter the real value of any endogenous variable in the long run. In the Appendix, we show that the system (in the nontraded goods case) is stable regardless of the sign of $1 - m - m^*$. Identical results hold for the traded goods case.

II. Devaluation and Nontraded Goods

A. *The Model*

In this section we investigate the role of nontraded goods in a devaluation. We continue to assume that each country produces two goods, but we impose the additional condition that transport costs prevent trade in Q_2 and Q_2^*. As in Section I the production-possibility frontier for each country is assumed concave. Thus, the domestic supply of any good remains solely a function of the domestic relative price of that good. Of the first ten equations in Section I, only equations (2) and (9) need modification. Since the markets for the nontraded goods are independent, P_2 need not equal eP_2^*. Furthermore, two equilibrium conditions are needed to replace equation (9) since the market for nontraded goods must clear in each country. Thus, we replace equation (2) with

$$P_1 = eP_1^* \tag{2'}$$

In place of equation (9), the conditions for the nontraded goods markets to clear are

$$Q_2(\rho) = D_2(\rho, E/P_1) \tag{15}$$

$$Q_2^*(\rho^*) = D_2^*(\rho^*, E^*/P_1^*) \tag{16}$$

where $\rho\,(\rho^*) = P_2/P_1(P_2^*/P_1^*)$

By Walras' Law, if equations (15), (16), and (7) hold, the market for traded goods must be in equilibrium. Thus, to adapt the model from the traded goods case to the nontraded goods case, equation (9) and the commodity arbitrage condition for good 2 are replaced by the conditions that the market in each country (for good 2) must be in equilibrium.

While equation (10) represents the trade balance for both the traded and nontraded goods cases, it is now more convenient to work with the real balance of trade measured in terms of the traded good (i.e., (B/P_1)). In the case of two traded goods, the meaning of the real balance of trade is somewhat ambiguous for it is possible to measure this balance in terms of importables or exportables. As the presence of only one traded good removes this ambiguity, and as it is desirable to work with real—as opposed to nominal—variables, we consider the effects of a devaluation on the real trade balance (B/P_1). Dividing equation (10) by P_1, and substituting $\pi W + \pi^* eW^*$ for $P_1(s\tilde{Y} + s^*\tilde{Y}^*)$, the balance of trade in terms of the traded good is

$$B/P_1 = \pi\pi^*[k\tilde{Y}eW^* - k^*\tilde{Y}^*W] \cdot [\pi W + \pi^* eW^*]^{-1} \tag{17}$$

Equation (17), like equation (10), shows that a *U.S.* trade deficit is caused by an excess supply of wealth in the United States relative to the United Kingdom.

B. *The Effects of a Devaluation*

As in Section I, the impact effect of a devaluation depends upon its ability to redistribute wealth. If a devaluation of the dollar increases *U.K.* wealth relative to *U.S.* wealth $(1 - m - m^* > 0)$, real *U.K.* expenditures will rise while real *U.S.* expendi-

tures fall. From equations (15) and (16), it is seen that the relative prices of nontraded goods are positively related to real expenditures. Thus, if $1 - m - m^* > 0$, the relative price of the nontraded good will rise in the United Kingdom and fall in the United States. However, if $1 - m - m^* = 0$, no relative redistribution of wealth occurs so that real expenditures and relative prices are unaltered. Finally, if $1 - m - m^* < 0$, the devaluation redistributes wealth *towards* the devaluing country, increasing real expenditures and the relative price of the nontraded good in that country, thereby producing perverse results. The critical factor to keep in mind is that the changes in relative prices are the effect of the devaluation, and in no sense can this relative price change be said to be the cause of the improvement or worsening of the balance of trade. The determining factor of the impact effect of the devaluation will always be how the devaluation redistributes wealth

These results can be obtained formally by totally differentiating equations (7), (15), and (16). Substitute equations (5), (11), and (12) into the above three to yield[9]

$$(18)\quad \frac{dP_1}{de}\frac{e}{P_1} = \left[\frac{Q_2Q_2^*}{\Delta}\right][\pi m W(\epsilon_2 - \eta_2)(\epsilon_2^* - \eta_2^* + s^*C_2^*) + \pi^*(1 - m^*)(eW^*)(\epsilon_2^* - \eta_2^*)(\epsilon_2 - \eta_2 + sC_2)] > 0 \quad \text{if } m > 0 \text{ or } m^* < 1$$

$$(19)\quad \frac{dp}{de}\frac{e}{p} = -C_2Q_2^*\pi\pi^* W(eW^*)(\epsilon_2^* - \eta_2^*)(1 - m - m^*)/[\Delta P_2] \lesseqgtr 0 \quad \text{as } 1 - m - m^* \gtreqless 0$$

$$(20)\quad \frac{dp^*}{de}\frac{e}{p^*} = C_2^* Q_2\pi\pi^* W(eW^*)(\epsilon_2 - \eta_2)(1 - m - m^*)/[\Delta P_2^*e] \gtreqless 0 \quad \text{as } 1 - m - m^* \gtreqless 0$$

where ϵ_2 (ϵ_2^*) = price elasticity of supply of good 2: $\epsilon_2, \epsilon_2^* > 0$

η_2 (η_2^*) = income compensated price elasticity of demand for good 2: $\eta_2, \eta_2^* < 0$

C_2 (C_2^*) = marginal propensity to consume good 2: $C_2, C_2^* > 0$

and

$$(21)\quad \Delta = Q_2Q_2^*[\pi^*eW^*(\epsilon_2^* - \eta_2^*)(\epsilon_2 - \eta_2 + sC_2) + \pi W(\epsilon_2 - \eta_2)\cdot(\epsilon_2^* - \eta_2^* + s^*C_2^*)] > 0$$

As previously argued, the devaluation affects relative prices only if it causes a wealth transfer $(m + m^* \neq 1)$. In particular, the relative price of the nontraded good decreases in the devaluing country (assuming both goods are normal) only if $m + m^* < 1$. Dornbusch's results hold since he assumes $m = m^* = 0$.

The impact of the devaluation on the balance of trade in terms of the traded good is found by differentiating (17), and substituting in for (18) (20):

$$(22)\quad \frac{d(B/P_1)}{de} = \left[\frac{\pi\pi^*}{\pi W + \pi^*eW^*}\right]\cdot\left[\frac{WW^*(1 - m - M^*)}{P_1} + kQ_2(eW^*)\frac{dp}{de} - k^*Q_2^*W\frac{dp^*}{de}\right]$$

where $sgn\ (dp^*/de) = -sgn\ (dp/de) = sgn\,(1 - m - m^*)$

Substituting for (dp/de), (dp^*/de) from (19) and (20):

$$(23)\quad \frac{d(B/P_1)}{de} = [\pi\pi^* WW^*(\epsilon_2 - \eta_2)\cdot(\epsilon_2^* - \eta_2^*)(1 - m - m^*)] \div P_1[\pi^*eW^*(\epsilon_2^* - \eta_2^*)(\epsilon_2 - \eta_2 + sC_2) + \pi W(\epsilon_2 - \eta_2)(\epsilon_2^* - \eta_2^* + s^*C_2^*)]$$

For $m = m^* = 0$, (23) is equivalent to the result derived in Dornbusch (his equation (27)).

First, from (23) we see that a devalua-

[9]The actual derivation is omitted in order to save space.

tion will improve the real balance of trade if and only if $(1 - m - m^*) > 0$. Thus, the necessary and sufficient condition for the devaluation to work in the nontraded good case (and in the traded good case when the terms of trade effect is ignored) is that it effectively transfers wealth away from the devaluing country.[10]

Next consider the role of changes in the relative prices of nontraded goods. From equations (19) and (20), it is seen that if the two relative prices of nontraded goods change $(1 - m - m^* \neq 0)$, they must move in opposite directions. This result follows from equations (15) and (16) in which the relative prices of nontraded goods are positively related to real expenditures. As real expenditures must rise in one country and fall in the other, the two relative prices move in opposite directions. Since the impact effect of a devaluation on real expenditures depends solely on $1 - m - m^*$, the direction of the changes in the relative prices of nontraded goods in contrast to the traded goods case depend solely upon the sign of $1 - m - m^*$. As the change in relative prices *is caused by changes in real expenditures*, it cannot be claimed that it is the change in relative prices which translates changes in real expenditures (absorption) into changes in the balance of trade. Rather, it is the desired change in expenditures which acts to alter relative prices.[11]

Further, the changes in relative prices act to offset part of the impact effect of a devaluation. From equation (22), it is clear that the improvement in the *U.S.* trade balance is positively related to dp/de and negatively related to dp^*/de. However, if the devaluation is to be successful (i.e., if $1 - m - m^* > 0$), the relative price of the nontraded good will fall in the United States and rise in the United Kingdom. Alternatively, if $1 - m - m^* < 0$, the relative price change will mitigate any deterioration in the *U.S.* trade balance due to a devaluation of the dollar. In short, the greater the change in relative prices, the *less* effective is the devaluation. The underlying explanation for this is clear—the *U.S.* trade balance can only be improved by increasing saving in the United States and correspondingly decreasing saving in the United Kingdom. A decrease in real wealth in the devaluing nation and a corresponding increase in real wealth in the revaluing nation (when $1 - m - m^* > 0$) serves this purpose. A decrease in real income in the devaluing nation, and an increase in real income in the revaluing nation, act to increase saving in the revaluing nation and reduce saving in the devaluing nation. Again, it should be pointed out that this is not simply an index number problem. The nominal *U.S* trade balance can only be improved by increasing nominal *U.S* saving. The greater the reduction in the relative price of the *U.S.* nontraded good, the smaller will be the rise in nominal *U.S.* income and correspondingly the smaller the rise in nominal *U.S* saving.

From the discussion above, it follows that if $1 - m - m^* > 0$, those factors which act to mitigate the size of relative price changes will act to increase the efficacy of a devaluation. Specifically, equation (23) demonstrates that if $1 - m - m^* > 0$, then the efficacy of a devaluation is negatively related to the marginal propensities to consume the nontraded good (C_2 and C_2^*)

[10] It should be pointed out that the dollar value of the trade balance may worsen even if $1 - m - m^* > 0$, and the pound value may increase even if $1 - m - m^* < 0$. Also notice that the sign of dB/de refers only to the balance of trade and not to the balance of payments. The balance of payments will equal the change in the demand for dollar denominated assets. If m and m^* are constant, the change in the demand for dollar denominated assets is $(1 - m)DW + m^*eDW^*$. As $DW + eDW^* = 0$, the balance of payments can be represented by $(1 - m - m^*)DW = (1 - m - m^*)B$. As the sign of dB/de depends upon the sign of $(1 - m - m^*)$, a devaluation of the dollar will always act to improve the *U.S.* balance of payments, whether or not it improves the balance of trade.

[11] Note that the whole emphasis of the elasticities approach is to determine how relative price changes improve the balance of trade, thereby increasing income relative to absorption. Further, both Dornbusch and Ronald Jones and W. M. Corden, imply that a relative price change acts to improve the trade balance. We find that the relative price change acts to worsen the trade balance of the devaluing nation.

and positively related to the absolute values of the supply and demand elasticities of nontraded goods ($\epsilon_2, \epsilon_2^*, |\eta_2|, |\eta_2^*|$). As expenditures fall (rise) in the devaluing (revaluing) nation, the greater will be the reduction (increase) in the price of the nontraded good if the marginal propensity to consume the nontraded good is large. Obviously, large supply and demand elasticities mitigate price changes so that the greater these elasticities, the greater the efficacy of a devaluation.

C. *Long-Run Effects of a Devaluation*

The discussion above analyzes the impact effect of a devaluation, and as such provides little information concerning the long-run effects of a devaluation or the stability of the system. In this section we first demonstrate that a devaluation does not alter the long-run equilibrium values of the real variables in the system. In the Appendix, we show that a fixed exchange rate system is stable regardless of the magnitude of $1 - m - m^*$.

Long-run equilibrium is obtained by setting the time derivatives $DW = DW^*$ equal to zero. Substituting these conditions into equations (5), (15), and (16):

$$Q_2(\rho) = D_2(\rho, \tilde{Y}(\rho)) \tag{24}$$

$$Q_2^*(\rho^*) = D_2^*(\rho^*, \tilde{Y}^*(\rho^*)) \tag{25}$$

$$W = kP_1\tilde{Y}(\rho) \tag{26}$$

$$eW^* = k^*P_1\tilde{Y}^*(\rho^*) \tag{27}$$

Equations (24)–(27), plus the condition that $W + eW^*$ is constant (as governments do not pursue an active monetary policy), determine the long-run values of ρ, ρ^*, W, W^*, and P_1. The nature of the steady-state solution of the model is that commodity markets are in equilibrium at each point in time. In addition, desired and actual saving in each of the countries are equal to zero. Given the equilibrium values of W and W^*, the amount of foreign assets held by *U.S.* (*U.K.*) residents is mW (mW^*).[12] With no net saving in either country, equation (6) indicates that the real balance of trade is equal to zero.

On inspection, it is immediately seen that equation (24) alone determines ρ and equation (25) alone determines ρ^*. Thus, long-run relative prices, real income, and commodity outputs are invariant to changes in the exchange rate or the outstanding supplies of dollar and pound denominated assets. Holding ρ and ρ^* constant, equations (26) and (27) can be used to solve for dP_1/de, dW/de, and $d(eW^*)/de$ once it is recognized that

$$\frac{dW}{de} + \frac{d(eW^*)}{de} = \bar{P} \tag{28}$$

where $\bar{P}$ ($\bar{D}$) = the total amount of assets which are denominated in pounds (dollars) when the devaluation takes place.[13]

Totally differentiating equations (26) and (27) with respect to the exchange rate, and using equation (28):

$$\frac{1}{P_1}\frac{dP_1}{de} = \frac{\bar{P}}{D + eP} \tag{29}$$

$$\frac{1}{W}\frac{dW}{de} = \frac{\bar{P}}{\bar{D} + e\bar{P}}$$

so that $d(W/P_1)/de = 0$. Given that $d(eW^*)/de = P - dW/de$, it follows that $d(eW^*/P_1)/de = 0$.

Thus, the exchange rate is neutral in the sense that the long-run magnitudes of all the real variables in the system are invariant with respect to a change in the exchange rate. In the Appendix, we demonstrate that the system has one characteristic root which

[12]The equilibrium conditions for the desired composition of assets to equal actual asset composition are

$$W_\$ = (1 - m)W + m^*eW^*$$

$$eW_P = mW + (1 - m^*)eW^*$$

where $W_\$$ = supply of dollar denominated assets
W_P = supply of pound denominated assets

[13]We assume that the central banks react passively in response to a deficit or surplus in that they neither sterilize nor accommodate the balance of payments. Further, central banks undertake no net wealth creation (except via devaluation), but stabilize exchange prices by making asset supplies perfectly elastic at the desired exchange rate.

is negative regardless of the value of $1 - m - m^*$. Thus, the equilibrium is stable and the approach to equilibrium is direct. Since the system has a single characteristic root, if the impact effect of a devaluation is to improve the real balance of trade, the devaluation will also act to increase the cumulative sum of the real trade balance. Consequently, a devaluation will serve to increase the central bank's holdings of foreign reserves (if $1 - m - m^* > 0$), even though the devaluation has no long-run effects on outputs and relative prices.

The discussion above indicates that a devaluation does not work through changing the long-run values of the real variables of the system, but rather through altering the time path of the system. In the short-run version of the model (Section IIA and IIB), desired and actual asset accumulations are equal, but desired portfolio size is not equal to actual portfolio size. In an attempt to equilibrate desired and actual wealth holdings, equation (5) indicates that individuals will save or dissave. As total world saving equals zero (since we assume total asset supplies are fixed) prices adjust such that if one nation has an excess demand for wealth, the other has an excess supply. The nation with an excess demand (supply) of wealth will experience a balance-of-trade surplus (deficit). The surplus (deficit), representing an increase (decrease) in the stock of wealth, serves to equilibrate desired and actual wealth holdings. The trade balance, then, is a temporary phenomenon which will persist until desired and actual wealth holdings are equal. Although the balance of trade is self-correcting (since the system is stable), the monetary authorities may desire to change the size of a deficit or surplus. As an exchange rate change does not alter the long-run equilibrium values of the real variables in the system, a successful devaluation must act to increase desired saving in the devaluing nation. We have shown that a devaluation will increase saving in the devaluing nation if $1 - m - m^* > 0$; the devaluation will redistribute wealth away from the devaluing nation towards the revaluing nation. It is this redistribution effect which determines whether saving will increase or decrease in the devaluing nation. Whether or not $1 - m - m^* > 0$, after the exchange rate change, the system will approach long-run equilibrium in which the trade balance is zero. The devaluation, then, only moves the system closer to, or further from, long-run equilibrium.

III. Conclusions

The approach taken in this paper is that balance-of-trade disequilibrium is caused by wealth imbalances between nations. Therefore, the primary effects of a devaluation must act to create an excess demand for wealth in the devaluing nation and an excess supply of wealth in the revaluing nation. To the extent that individuals view domestic and foreign assets as substitutes, the redistributive effects of a devaluation will be reduced so that exchange rate policies will be of little value in correcting deficits or surpluses in the trade accounts. As the volume of trade is one of the major determinants of the degree of asset substitutability between equally risky assets, we would expect that the degree of openness of an economy and the efficacy of a devaluation within that economy are negatively related.

We have also examined the effects of relative price changes on the efficacy of a devaluation. Relative price changes will act to alter a trade balance to the extent that they change the demand for wealth. However, in the case of nontraded goods, relative prices always change in such a way that they offset part of the effects of the devaluation. In the case in which both goods are traded, the terms-of-trade effect may act to either increase or decrease the trade balance. In both cases, however, the relative price change cannot be said to be the cause of the change in the trade balance. Rather it is the effects of changes in desired expenditures which induce the change in relative prices.

APPENDIX

In this Appendix we consider the stability conditions for fixed and flexible exchange rate regimes in the nontraded good case. We demonstrate that either exchange rate regime is stable regardless of the sign of $1 - m - m^*$.

Equations (15) and (16) can be written as

$$\text{(A1)} \quad Q_2(\rho) = D_2\left(\rho, \tilde{Y}(\rho) - \frac{1}{P_1}(DW)\right)$$

as $E/P_1 = \tilde{Y} - (1/P_1)DW$ and $\tilde{Y} = Y(\rho)$

$$\text{(A2)} \quad Q_2^*(\rho^*) = D_2^*\left(\rho^*, \tilde{Y}^*(\rho^*) - \frac{e}{P_1}DW^*\right)$$

as $E^*/P_1^* = \tilde{Y}^* - (1/P_1^*)DW^*$ and $P_1/e = P_1^*$.

From equations (5), (7), and (8),

$$\text{(A3)} \quad DW = kP_1\tilde{Y}(\rho) - W$$

where the adjustment parameters π and π^* have been set equal to unity

$$\text{(A4)} \quad eDW^* = k^*P_1\tilde{Y}^*(\rho^*) - eW^*$$

The U.S and U.K. residents hold both dollar and pound denominated assets in their portfolios. The dollar value of their portfolios can be represented by

$$\text{(A5)} \quad W = W^{\$} + eW^{P}$$

$$\text{(A6)} \quad eW^* = W^{*\$} + eW^{*P}$$

where $W(eW^*)$ = dollar value of wealth held by U.S. (U.K.) residents

$W^{\$}(W^{*\$})$ = dollar denominated assets held by U.S. (U.K.) residents

$W^{P}(W^{*P})$ = pound denominated assets held by U.S. (U.K.) residents

As in the text, it is assumed

$$\text{(A7)} \quad W^{\$} = (1 - m)W$$

$$\text{(A8)} \quad W^{*P} = (1 - m^*)W^*$$

Asset market equilibrium requires that the demand for dollar (pound) denominated assets equals the amount outstanding:

$$\text{(A9)} \quad W_{\$} = W^{\$} + W^{*\$}$$

$$\text{(A10)} \quad W_{P} = W^{P} + W^{*P}$$

where $W_{\$}(W_P)$ = outstanding amount of dollar (pound) denominated assets

The system can be simplified by combining equations (A5)–(A10) to yield

$$\text{(A5}'\text{)} \quad W_{\$} = (1 - m)W + m^*eW^*$$

$$\text{(A6}'\text{)} \quad W_P = m\frac{W}{e} + (1 - m^*)W^*$$

Equations (A1)–(A4), (A5′), and (A6′) represent six independent equations containing eight unknowns: ρ, ρ^*, W, W^*, P_1, e, $W_{\$}$, and W_P. Under flexible exchange rates, $W_{\$}$ and W_P can be treated as constants since central banks do not attempt to alter asset supplies. With a fixed exchange rate, e is constant; and as it is assumed that governments only alter money supplies in response to the balance of payments: $DW_{\$} = -eDW_P$. Thus, when it is known whether the exchange rate is fixed or flexible, two unknowns are eliminated from the system.

The dynamic model which we postulate is quite different from that used by Aghevli and Borts. The nature of our model is such that commodity markets are in equilibrium at each point in time. Given the existing stocks of wealth within countries, desired and actual portfolio compositions are also equal. The dynamic nature of our model is due to the fact that desired asset accumulations change over time. Asset accumulation acts to partially eliminate the discrepancy between desired and actual wealth, thereby reducing desired saving.

A. *Fixed Exchange Rates*

Equations (A1)–(A4) and the equation $DW = -eDW^*$ constitute five equations containing five unknowns. Substitute equations (A3) and (A4) into equations (A1) and (A2). Then linearize the resulting two equations as well as equations (A3) and (A4) around the point of long-run equi-

$$
(A11)\quad \begin{bmatrix}
\frac{Q_2}{\rho}(\epsilon_2 - \eta_2 + sC_2) & 0 & \frac{1}{\rho}\left(\frac{C_2 W}{P_1^2}\right) & \frac{-1}{\rho}\left(\frac{C_2}{P_1}\right) & 0 \\
0 & \frac{Q_2^*}{\rho^*}(\epsilon_2^* - \eta_2^* + s^* C_2^*) & \frac{1}{\rho^*}\left(\frac{C_2^* e W^*}{P_1^2}\right) & 0 & \frac{1}{\rho^*}\left(\frac{-C_2^*}{P_1}\right) \\
0 & 0 & 0 & D & D \\
sP_1 Q_2 & 0 & s\tilde{Y} & -(1+D) & 0 \\
0 & s^* P_1 Q_2^* & s^* \tilde{Y}^* & 0 & -(1+D)
\end{bmatrix}
\begin{bmatrix} \rho \\ \rho^* \\ P_1 \\ W \\ eW^* \end{bmatrix} = K_0
$$

librium. The resulting four equations and the relation $DW = -eDW^*$ can be represented by (A11), where $K_0 = 5 \times 1$ column vector of constants, and $\pi = \pi^* = 1$ so that $s = k$ and $s^* = k^*$

The characteristic equation takes the form $D(a_1 D + a_2)$. Thus, the system has a single nonzero characteristic root equal to $-a_2/a_1$. The system will be stable if a_1 and a_2 are of the same sign Solving for a_1 and a_2:

$$a_1 = \frac{\Delta}{\rho\rho^* P_1}, \quad \text{where } \Delta \text{ is defined in equation (21)}, a_1 > 0$$

$$a_2 = \frac{Q_1 Q_2}{\rho\rho^*}(\epsilon_2 - \eta_2)(\epsilon_2^* - \eta_2^*) \cdot (k\tilde{Y} + k^*\tilde{Y}^*) > 0$$

As both a_1 and a_2 are unambiguously positive, the system is stable regardless of the sign of $1 - m - m^*$

B. *Flexible Exchange Rates*

Equations (A7) (A4), (A5′), and (A6′) represent six equations in six unknowns. Again, substitute equations (A3) and (A4) into equations (A1) and (A2). Linearizing the resulting two equations, plus equations (A3), (A4), (A5′), and (A6′) around the point of long-run equilibrium yields (A12), where $K_1 = 6 \times 1$ column vector of constants.

As the differential operator appears twice in position a_{35} and a_{46}—the system will have two nonzero characteristic roots at most, that is, the characteristic equation will take the form $\alpha_1 D^2 + \alpha_2 D + \alpha_3 = 0$. On setting the determinant of the coefficient matrix equal to zero, the characteristic equation actually takes the form $a_4 D + a_5 = 0$. Thus, the system has only one nonzero root which is equal to $-a_5/a_4$. The system will be stable if a_5 and a_4 are of the same sign.

On setting the initial values of e, P_1 and P_1^* equal to unity, we find:

$$a_5 = \frac{Q_2 Q_2^*}{\rho\rho^*} C_2 C_2^* (\epsilon_2 - \eta_2)(\epsilon_2^* - \eta_2^*) \cdot m^* W^* [W^*(1 - m^*) + mW]$$

$$a_4 = \frac{Q_2 Q_2^*}{\rho\rho^*} C_2 C_2^* \{(\epsilon_2 - \eta_2 + sC_2) k^* \tilde{Y}^* \cdot (\epsilon_2^* - \eta_2^*) m^* [(1 - m^*) W^* + mW] + (\epsilon_2^* - \eta_2^* + s^* C_2^*) k \tilde{Y} (\epsilon_2 - \eta_2) \cdot m[(1 - m) W + m^* W^*]\}$$

$$
(A12)\quad -\begin{bmatrix}
\frac{Q_2}{\rho}[\epsilon_2 - \eta_2 + sC_2] & 0 & \frac{1}{\rho}\left(\frac{C_2 W}{P_1^2}\right) & 0 & -\frac{1}{\rho}\left(\frac{C_2}{P_1}\right) & 0 \\
0 & \frac{Q_2^*}{\rho^*}[\epsilon_2^* - \eta_2^* + s^* C_2^*] & \frac{1}{\rho^*}\left(C_2^* e \frac{W^*}{P_1^2}\right) & \frac{-1}{\rho^*}\left(C_2^* \frac{W^*}{P_1}\right) & 0 & -\frac{1}{\rho^*} C_2^* \frac{e}{P_1} \\
-sP_1 Q_2 & 0 & s\tilde{Y} & 0 & (1+D) & 0 \\
0 & -s^* \frac{P_1}{e} Q_2^* & s^* \frac{\tilde{Y}^*}{e} & S^* P_1^* \frac{\tilde{Y}^*}{2} & 0 & (1+D) \\
0 & 0 & 0 & m^* W^* & (1-m) & m^* e \\
0 & 0 & 0 & -\frac{mW}{e^2} & \frac{m}{e} & (1-m^*)
\end{bmatrix}
\begin{bmatrix} \rho \\ \rho^* \\ P_1 \\ e \\ W \\ W^* \end{bmatrix} = K_1
$$

Since m and m^* are both positive fractions, a_4 and a_5 are both positive.

REFERENCES

B. B. Aghevli and G. H. Borts, "The Stability of Equilibrium of the Balance of Payments Under a Fixed Exchange Rate," *J. Int. Econ.*, Feb. 1973, *3*, 1-20.

R. Dornbusch, "Devaluation, Money, and Nontraded Goods," *Amer Econ Rev.*, Dec. 1973, *63*, 871-80.

W. Enders, "Portfolio Balance and Exchange Rate Stability," *J. Money, Credit, Banking*, Aug. 1977, *9*, 491-99.

Jacob A. Frenkel and Harry G. Johnson, *The Monetary Approach to the Balance of Payments*, Toronto 1976.

——— and C. Rodriguez, "Portfolio Equilibrium and the Balance of Payments," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674-94.

A. Harberger, "Currency Depreciation, Income, and the Balance of Trade," *J Polit. Econ.*, Feb. 1950, *53*, 47-60.

R. Jones and W. M. Corden, "Devaluation, Non-Flexible Prices, and the Trade Balance for a Small Country," *Can. J. Econ.*, Feb. 1976, *9*, 150-61.

H. G. Johnson, "The Monetary Approach to the Balance of Payments," in Jacob Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, Toronto 1976.

S. Laursen and L. Metzler, "Flexible Exchange Rates and the Theory of Employment," *Rev. Econ. Statist.*, Nov. 1950, *32*, 281-99.

D. Mathieson, "Portfolio Disequilibrium, the Speed of Adjustment and the Balance of Payments," unpublished paper, Columbia Univ. 1974.

R. McKinnon, "Portfolio Balance and International Payments Adjustment," in Robert Mundell and Alexander Swoboda, eds., *Monetary Problems of the International Economy*, Chicago 1969.

S. C. Tsiang, "The Role of Money in Trade-Balance Stability: Synthesis of the Elasticity and Absorption Approaches," *Amer. Econ. Rev.*, Dec. 1961, *51*, 912-36.

A Theory of Pricing under Decreasing Costs

By J. SORENSON, J. TSCHIRHART, AND A. WHINSTON*

Selecting appropriate pricing policies is a central problem in economics. The subject has a long history especially with regard to marginal-cost pricing as an appropriate policy for decreasing cost industries. In these industries marginal-cost pricing implies deficits and other associated problems. The arguments largely concerned with the welfare economic aspects of marginal-cost pricing have been summarized by Nancy Ruggles.

The direction of this paper is somewhat different. Our approach takes the producer and the potential consumer as a unit and attempts to delineate the class of viable pricing schemes. A potential consumer is characterized by a demand curve for the product or service produced, while the producer is characterized by the cost of production. We assume that only one good is produced, but allow for different demand curves among the consumers.

Given the description of the problem, we consider the benefits that are available to both the producer and the consumers of the product. For the producer, the benefit is the difference between total revenue received by whatever pricing scheme is used and the total cost of production. For each consumer, the benefit is the difference between the value of his total utility from the product and the total charge. If under a particular charge formula the benefit to the producer is negative, then we assume that this would be unacceptable from the viewpoint of the producer. This would rule out marginal-cost pricing as an acceptable scheme. Likewise, each consumer must receive a positive net benefit to be willing to accept a particular solution. A pricing scheme should also be efficient in that total benefits received by all parties are at a maximum. Given these boundary conditions, we inquire into whether there exists any pricing scheme, whether it be a uniform price per unit or a highly non-linear price schedule that will be acceptable to all parties. By acceptance we mean that no party will find it preferable to withdraw from either producing or consuming the good under the arrangements proposed We shall see that this last point is somewhat subtle in that it does not necessarily mean giving up either producing or consuming the good, but that other arrangements are preferable that exclude some consumers.

Our first result shows that there do exist pricing rules which lead to a stable arrangement. In fact, there are many different rules which are acceptable from this point of view. The rules differ in their distributional impact Next we analyze specific pricing rules that have been proposed. We determine whether a proposed rule is stable in a sense to be defined, and within our framework, analyze the distributional aspects of each rule. Finally, since our model is cast in terms of game theory, we consider several game-theoretic solutions as pricing rules.

A game-theory model derives naturally from the subject of this paper. Potential players of the game (i.e., consumers) find it beneficial to aggregate their demands since costs are decreasing. The total benefits grow as coalition sizes increase, and the game becomes one of determining how the benefits are to be distributed. Selecting a method of distributing benefits is tantamount to selecting a price rule The problem is similar to one discussed by James Buchanan (1965) and Mark Pauly in relation to the theory of clubs where the number of consumers is a

*Professor of mathematics, Valparaiso University, assistant professor of economics, University of Wyoming, and professor of economics, management and computer science, Purdue University, respectively. This research was supported in part by National Science Foundation Grant Number ENG 75-07845. We are indebted to the managing editor and a referee for valuable suggestions on an earlier draft.

variable. By decreasing marginal cost, the optimum club size in this paper consists of all potential consumers. However, this does not assure stability. It is necessary that every subset of consumers within the total set is satisfied with the pricing arrangements. If this is not so, the total set is unstable and may disband. We will show that traditional pricing schemes (for example, two-part tariffs) do not guarantee stable pricing arrangements.

I. The Pricing Game

A. *The Characteristic Function*

We use consumer's surplus as a measure of benefits received, and assume that income elasticity is zero so that the ordinary and compensated demand curves coincide.[1] This assumption is not unreasonable for goods such as water, sewage, or even electricity, since the total amount spent on these goods is usually a small fraction of income. The advantage of assuming zero income elasticity is that for a particular quantity demanded by a consumer, we can vary within limits the total charge to the consumer through a lump sum tariff without changing the quantity.[2] This allows us to focus attention on the distributional problem.

Let $N = \{1, \ldots, n\}$ be the set of players in the game where index 1 denotes the entrepreneur and indices $2, \ldots, n$ denote the consumers. Consumer i's demand function is given by $\Psi_i(q_i)$ which is assumed to be monotonically decreasing over the quantity interval $0 \leq q_i \leq q_i^*$. The amount consumer i is willing to pay for q_i rather than go without is

$$A_i = \int_0^{q_i} \Psi_i(t)dt$$

For any subset of consumers $S \subset N - \{1\}$, we have an aggregate demand correspondence $\Psi_S(q)$, where $q = \Sigma_S q_i$, obtained by horizontally adding the demands of each consumer in S. The amount that any subset of consumers is willing to pay for q rather than go without depends on how q is allocated among the consumers. Let $A_S(q)$ be this amount defined by the maximum $\{\Sigma_S A_i(q_i): q_i \geq 0, \Sigma_S q_i = q\}$. Thus $A_S(q)$ is the maximum benefit that can be attained by allocating q among the various consumers. The maximum is attained when q is allocated so that $q_i = \Psi_i^{-1}(\Psi_S(q))$, since any other allocation can be improved upon by increasing (decreasing) q_i when $q_i < (>)$ $\Psi_i^{-1}(\Psi_S(q))$. Therefore, we have

$$A_S(q) = \int_0^q \Psi_S(t)dt = \sum_S \int_0^{q_i} \Psi_i(t)dt$$

where q is allocated in the above manner.

For the entrepreneur, the marginal cost of production is $c(q)$ which is assumed to be nonincreasing and continuous over the interval $[0, q(N)]$, where $q(N) \leq \Sigma_N q_i^*$ is the total output defined below. This rules out increasing marginal cost over the relevant range of production. The total cost of producing q is

$$C(q) = \int_0^q c(t)dt$$

Before defining the characteristic function, we make the following observations. First, any coalition which does not contain the entrepreneur cannot produce the good. Consequently, the maximum payoff that this coalition can expect to obtain on its own is zero. This essential role played by the entrepreneur may be due to the lack of technical skills, entrepreneurship, coordination, etc. on the part of the consumers. Conversely, the coalition that con-

[1] A good exposition on this subject may be found in Don Patinkin. The point here is that if income elasticity is zero, the demand curve in the quantity-price plane is invariant with respect to income changes. For this to occur, it is necessary that the consumer's indifference curves are parallel, that is, at each quantity, the slopes of all indifference curves are equal. A utility function that will result in zero income elasticity for good y is $U(x, y) = x + ln\ y$

[2] If we confined our analysis to uniform prices, then income elasticity would not play an important role. However, as Buchanan observes, considering only uniform prices is very restrictive, "... quantity discounts, quantity premiums, block tariffs generally are both institutionally feasible and empirically observable" (1966, p. 465)

tains no consumers, the entrepreneur alone, also can expect to gain zero at most. Second, the cost of production must be born solely by the players in the game. This rules out the possibility of covering deficits through taxation on the general public. The definition of the characteristic function v is as follows:

$$(1)\quad v(S) = \begin{cases} 0 & \text{for } 1 \notin S \text{ or } S = \{1\} \\ \max\{A_S(q) - C(q), q \geq 0\} & \\ & \text{for } 1 \in S \text{ and } S \neq \{1\} \end{cases}$$

for $q = \Sigma_S q_i$ and for all $S \subseteq N$. For $1 \in S$ and $S \neq \{1\}$, the quantity that maximizes $v(S)$ is denoted by $q(S)$, and the quantity received by consumer $i \in S$ is denoted by $q_i(S)$. For convenience, $q_1(S) = 0$ and $A_1(q_1(S)) = 0$ for all $S \subseteq N$. Since A_S and C are continuous, this maximum is attained either at $q(S) = 0$ or the point(s) for which $\Psi_S(q(S)) = c(q(S))$. If there is more than one point where the maximum is attained, we will take the minimum for $q(S)$. The maximum is zero and $q(S) = 0$ when $A_S(q) - C(q) \leq 0$ for all $q > 0$. This result, of course, is well known. Maximum welfare is attained at that quantity $q(N)$ where demand is equal to marginal cost, and the quantity is allocated according to $q_i(N) = \Psi_i^{-1}(\Psi_N(q(N)))$.[3]

An important property of this game that will be used throughout is as follows:

PROPOSITION 1: *The characteristic function for the pricing game is convex.*

The proof of this proposition appears in the Appendix. The concept of convexity is analogous to increasing returns in economics. A characteristic function is convex if and only if

$$(2)\quad v(S \cup k) - v(S) \leq v(T \cup k) - v(T)$$

for all $S \subseteq T \subseteq N - \{k\}$

Therefore, the incremental benefit generated by a player upon joining coalition T is at least as much as that for joining coalition S if $S \subset T$. By convexity, the core exists and tends to be relatively "large." For a complete description of the relationships between cores and convex games, see Lloyd Shapley (1971).

B. *Payoffs, Charges, and a Numerical Example*

The payoff or net benefit x_i to consumer i in the pricing game is measured by the total benefit derived from $q_i(N)$ minus the charge, $g(i)$, levied on consumer i. If $g(i)$ is equal to the total benefit, then the payoff is zero and consumer i is neither better nor worse off than if he had never participated. The payoff to the entrepreneur is the profit π, which measures total revenue minus total cost. These payoffs are written as follows:

$$(3)\quad x_1 = \pi = \sum_2^n g(i) - C(q(N))$$
$$x_i = A_i(q_i(N)) - g(i), i = 2, \ldots, n$$

In game theory terminology, the payoff vector $x = (x_1, x_2, \ldots, x_n)$ is an imputation if it satisfies individual rationality.

$$x_1 \geq v(1) = 0 \rightarrow \sum_2^n g(i) \geq C(q(N))$$
$$x_i \geq v(i) = 0 \rightarrow A_i(q_i(N)) \geq g(i), \quad i = 2, \ldots n,$$

and rationality on the part of the grand coalition,

$$\sum_N x_i = v(N) = A_N(q(N)) - C(q(N))$$

Note that individual rationality on the part of the entrepreneur requires that total cost

[3] A special case of this game can be found in papers by Dermot Gatley, Stephen Littlechild and G. Owen, and Edna Loehman and Andrew Whinston. These authors assume $q_i(N) = q_i(S)$ for all $i \in S \subseteq N - \{1\}$ so that demands are perfectly inelastic. Under these conditions it can be shown that the characteristic function is strategically equivalent to $v(S) = \Sigma_S C(q_i) - C(\Sigma_S q_i)$. (See R. Duncan Luce and Howard Raiffa for a definition of strategic equivalence.) Thus evaluation of utility is unnecessary. A property of this special case that does not hold in general is that if consumer j generates more incremental value when he joins a coalition than consumer k generates, then consumer j generates more incremental value than k upon joining any coalition.

be covered. Hence, only full cost-pricing rules will result in imputations. The core is a subset of the imputations, which requires rationality on the part of all coalitions. Therefore, x is in the core if for all $S \subseteq N$,

$$\sum_S x_i \geq v(S) \tag{4}$$

The following is a numerical example with two consumers that illustrates the above concepts. The individual and aggregate demand curves are

$$\Psi_2(q_2) = -\frac{2}{3}q_2 + 30$$

$$\Psi_3(q_3) = -2q_3 + 30$$

$$\Psi_N(q_N) = -.5q_N + 30$$

and the entrepreneur's total and marginal-cost functions in the relevant quantity region are

$$C(q) = -\frac{1}{16}q^2 + 15q$$

$$c(q) = -\frac{1}{8}q + 15$$

With this information, (1) is determined as follows:

$$v(1,2) = A_{12}(27.7) - C(27.7) = 207.7$$

$$v(1,3) = A_{13}(8.0) - C(8.0) = 60.0$$

$$v(N) = A_N(40.0) - C(40.0)$$

$$= A_2(30.0) + A_3(10.0) - C(40.0) = 300.0$$

and $v(S) = 0$ for all other $S \subset N$.

The fundamental triangle (the set of all imputations) for this example is constructed in Figure 1. The payoff to player i is measured along axis x_i and the core is delimited by points *abcde*.

II. Traditional Pricing Methods

The following pricing methods are well known. The advantage of incorporating them in a game-theoretic setting is that it allows a clear representation of the significant similarities, differences, and the stability properties.

FIGURE 1

A. Discriminating Two-Part Tariff-Profit Maximization

This pricing method is ably discussed in Walter Oi's Disneyland Dilemma. The entrepreneur maximizes profit given by

$$\sum_2^n (\Psi_i(q_i)q_i + T_i) - C(q)$$

where T_i is a nonnegative license fee for consumer i. For a particular quantity, the consumer has a consumer's surplus of $A_i - \Psi_i(q_i)q_i$. When a license fee is charged, it is analogous to decreasing the consumer's income. This drives the consumer to a lower indifference curve: but by zero income elasticity the consumer does not change the quantity purchased. There is a bound, however, on the T_i. If T_i exceeds consumer's surplus, then quantity is zero because the total charge exceeds the derived benefit. Therefore, the maximization is subject to the constraints

$$T_i \leq A_i - \Psi_i(q_i)q_i \qquad \text{for } i = 2, \ldots, n \tag{5}$$

The maximum π^* is attained where price equals marginal cost, and (5) is satisfied by an equality for $i = 2, \ldots, n$. Thus, the entrepreneur captures the entire consumer's surplus by establishing a personalized license fee for each consumer. The resulting allocation is $x_1 = \pi^*$ and $x_2 = x_3 = 0$, and is

given by point e in Figure 1. From the standpoint of fairness, there are obvious objections one could raise in opposition to the core point e.

B. *Discriminating Two-Part Tariff–Welfare Maximization*

With this method, the consumer's welfare is maximized, while the entrepreneur's profit is constrained to a nonnegative quantity. Using the same notation, we have

$$(6)\quad \max W = \sum_2^n (A_i - \Psi_i(q_i)q_i - T_i)$$

subject to (5) and

$$(7)\quad \sum_2^n (\Psi_i(q_i)q_i + T_i) - C(q) = \bar{\pi}$$

Constraint (7) requires a fixed profit while (5) ensures that no consumer pays more for the good than it is worth to him.

It can easily be shown that maximum profit is attained where $\Psi_i(q_i(N)) = \Psi_N(q(N)) = c(q(N))$. For notational convenience, we let $p = \Psi_N(q(N))$ below. The allocation of the remaining benefit is subject to

$$\sum_2^n x_i = v(N) - \bar{\pi}$$

and $0 \leq x_i \leq A_i(q_i(N)) - pq_i(N)$ for $i = 2, \ldots, n$. However, to ensure a core allocation (4) is an additional restriction that must be imposed. Thus if C_R and F denote the core and the set of allocations generated by the optimization problem when $0 \leq \bar{\pi} \leq \pi^*$, then $C_R \subseteq F$. The implication is that if subcoalitions can strike up separate contracts with the entrepreneur, the set of charges levied on the consumers in the optimization problem may contain unstable allocations. If $x \in F$ but $x \notin C_R$, then at least one group of consumers can do better by dropping out of the grand coalition and bargaining separately with the entrepreneur.

This point can be illustrated using our example. Assuming a break-even constraint so that $\bar{\pi} = 0$, the set of allocations generated by the optimization problem are calculated by noting that $p = 10$, $q_2(N) = 30$, and $q_3(N) = 10$ so that constraints (5) and (7) yield

$$0 \leq T_2 \leq 300$$
$$0 \leq T_3 \leq 100$$
$$T_2 + T_3 = 100$$

This set of allocations is represented by line segment fh in Figure 1. Point f corresponds to $T_2 = 0$, $T_3 = 100$, and point h corresponds to $T_2 = 100$, $T_3 = 0$.

Along fh, only those points on line segment bc are in the core. This indicates that a more restrictive set of license fees are needed to ensure stability. To obtain these fees we note that b and c are extreme points of the core in a convex game, and are therefore represented by the payoff vectors $(v(1), v(N) - v(13), v(13) - v(1))$, and $(v(1), v(12) - v(1), v(N) - v(12))$, respectively (see Shapley (1971)). In turn, these payoffs and (5) and (7) yield

$$60 \leq T_2 \leq 92.3$$
$$7.7 \leq T_3 \leq 40$$
$$T_2 + T_3 = 100$$

The noncore allocations on segments fb and ch arise because the optimization problem does not account for opportunities available to subcoalitions. For instance, if allocation f were proposed, the net benefit to consumer 3 is $A_3(q_3(N)) - pq_3(N) - T_3 = 0$. Consumer 3 could improve upon this situation by striking a separate bargain with the entrepreneur and receiving as much as $v(1,3) = 60$.

Two more examples are shown in Figure 1 by line segments $f'h'$ and $f''h''$ corresponding to $W^* = 150$, $\pi = 150$ and $W^* = 50$, $\pi = 250$, respectively.

C. *Uniform Two-Part Tariff*

With a uniform two-part tariff $T_i = T$ for $i = 2, \ldots, n$. Since this adds another set of constraints to the welfare-optimization problem, the optimal solution must remain unchanged or decrease. A decrease occurs if some consumers abandon the market entirely because T is too large. In this case

$v(N)$ is not attained and a loss of total welfare results.

To illustrate this, we maximize (6) subject to (7). The only difference now is that T_i is replaced by T in both expressions. From the first-order conditions, $p = c(q(N))$ and

$$T = \frac{C(q) + \bar{\pi} - pq(N)}{n - 1} \tag{8}$$

The license fee is the portion of cost and profit not covered by per unit price, divided by the number of consumers. This solution satisfies

$$\sum_1^n x_i = v(N)$$

provided that

$$T \leq \min \{A_i - pq_i(N), i = 2, \ldots, n\} \tag{9}$$

Equality in (9) implies that $\bar{\pi}$ is the maximum attainable profit given that price equals marginal cost and all consumers are included in the market. However, this may not be the maximum profit available; excluding some consumers by violating (9), and/or deviating from marginal-cost pricing may yield a greater π but

$$\sum_1^n x_i < v(N)$$

Although this latter situation can potentially be changed to the advantage of everyone, the presence of the uniform tariff prohibits the change.

In terms of the example, the maximum attainable profit when all consumers are included in the market and $v(N)$ is realized requires from (9) $T = A_3 - pq_3(N) = 100$. Substituting this into (8) with $p = 10$ and $C(q) = 500$ yields $\bar{\pi} = 100$. Thus $x_1 = 100$ and from (3) with $g(i) = pq_i(N) + T$, $x_2 = 0$ and $x_3 = 200$. Unlike the discriminatory two-part tariff, a uniform two-part tariff yields a unique allocation for each $\bar{\pi}$. For $\bar{\pi} = 100$, this corresponds to point w' in Figure 1. Using similar calculations in the uniform two-part tariff case, we can obtain point w where $T = 50$, $x_1 = \bar{\pi} = 0$, $x_2 = 250$, and $x_3 = 50$; or any point on the line segment ww' along which $50 \leq T \leq 100$ and $0 \leq \bar{\pi} \leq 100$. Note that ww' includes allocations inside and outside the core; therefore, uniform tariff structures also have instability problems. In particular, any point on the segment connecting w and the intersection of ww' with ab will be unsatisfactory to coalition $\{1,3\}$ because of a combination of small profits and a relatively large license fee for consumer 3.

To maximize π in this example, it is necessary to exclude consumer 3 by setting $\Psi_S(q(S)) = c(q(S))$ and $T = A_2(q_2(S)) - \Psi_S(q(S))q_2(S)$ for $S = \{1,2\}$. We obtain $q_2(S) = 27.7$, $\Psi_S(q(S)) = 11.5$ and $T = 255.5$ so that $x_3 = x_2 = 0$ and $x_1 = 207.7 = v(1,2)$. This global profit maximum is a one-consumer one-producer case where the entire consumer's surplus is exhausted by the two-part tariff. Point u on the x_1 axis in Figure 1 represents this solution. The circle indicates that u is below (closer to the origin) the plane of the triangle. There are many points in the core which are superior in the sense that all players could be made better off.

III. Game-Theoretic Pricing Rules

Solutions that allocate the value of a game abound in the game-theory literature. The core itself is a type of solution. If it exists, it contains the set of outcomes that cannot be improved upon by any coalition; consequently, the definition of the core is easily defended as a minimum criteria for a viable solution. Since the core does exist in our pricing game, we can justifiably eliminate pricing methods which result in noncore outcomes. The results of the previous section indicate that some traditional methods may fall into this category. The pricing game is also convex so that the core is large, and there are many outcomes to choose from; unless one is willing to make ethical judgments, the problem of selecting a unique outcome remains. We are confronted with a distributional problem of selecting one Pareto optimum allocation from among many.

In what follows we consider two game-theoretic solutions that select a particular Pareto optimum and thereby assign unique

charges to each player. These solutions are the Shapley value and the nucleolus. In Section II, g took on the familiar form of a two-part tariff (i.e., $g(i) = pq_i + T_i$). The game-theoretic solutions yield a single lump sum charge to consumer i for quantity $q_i(N)$. As Oi suggests, however, this charge can always be reinterpreted as a two-part tariff. Consequently, the game-theoretic solutions are special cases of multipart tariff structures.

A. *The Shapley Value*

The Shapley value was introduced by Shapley (1953), and has been discussed as a cost allocation scheme by Martin Shubik, Littlechild and Owen, and Loehman and Whinston. By convexity the Shapley value is in the "center" of the core in the cost allocation game.[4] Using the characteristic function in our example and the Shapley value formula

$$x_i = \sum_{\substack{S \subset N \\ i \in S}} \frac{[(s-1)!(n-s)!]}{n!} \cdot [v(S) - v(S - \{i\})]$$

where s is the number of players in coalition S, the payoffs are $x_1 = 144.6$, $x_2 = 114.6$, and $x_3 = 40.8$. Point s^* in Figure 1 represents this allocation.

It is difficult to say whether this allocation is fair, because it is not clear how one compares the entrepreneur's payoff with the consumers' payoffs. In terms of profit, if $C(q(N))$ is the cost of capital then the return on investment is over 50 percent. The Shapley value awards this large payoff, because of the entrepreneur's asymmetric role in the game. Unlike the other players, the entrepreneur has no demand; yet without him all coalitions are powerless. Occupying this strategic position means that $1/n$th of the core extreme points coincide where the entrepreneur receives the full benefit (point e in Figure 1).[5] The Shapley value is an evenly weighted combination of the extreme points; consequently, the entrepreneur always receives $v(N)/n$ plus a weighted portion of the other extreme points.

Normally, an entrepreneur can not expect, nor do we observe 50 percent profits. However, this 50 percent is dependent on the particular example. If consumers' surpluses are less, then potential profit is less. There may also be other reasons why 50 percent is not attained in practice. We have assumed that coalitions without the entrepreneur are powerless by their inability to produce. We can relax this assumption by allowing consumer coalitions to produce subject to a cost function C_c, where $C_c(q) > C(q)$ for $q > 0$. Consumers can produce on their own but not as efficiently as an entrepreneur. Then if the entrepreneur insists on a profit that consumers find unacceptable, consumer co-ops may form and undermine his profit. The effect of consumer co-ops on the game is to alter the characteristic function. Instead of $v(S) = 0$ for $1 \notin S$, we would have

$$v(S) = \max_q \{A_S(q) - C_c(q), q \geq 0\}$$

This clearly weakens the entrepreneur's power and decreases his Shapley payoff. Other phenomena that may limit profits are competition and regulation.

B. *The Nucleolus*

A consumer may agree to maximize the payoff to the least well-off coalition if he feels he may be in that coalition.[6] A game-theoretic solution that embodies this type of payoff is the nucleolus as proposed by David Schmeidler. For each payoff vector $x \in R^n$, where $x_i \geq 0$, $i = 1, \ldots, n$ and

$$x(N) = \sum_{i \in N} x_i = v(N)$$

---

[4]Center refers to the center of gravity of the core vertices. See Shapley (1971).

[5]The core of a strictly convex game (marginal cost strictly decreasing) has $n!$ extreme points corresponding to the $n!$ possible orderings of the players. An extreme point corresponds to a payoff where every player receives the marginal benefit he generates as he enters the game. See Shapley (1971).

[6]There is an analogy here with John Rawls's proposal for just allocation.

Let $\theta(x)$ be a vector in R^{2^n}.[7] The components of $\theta(x)$ are $v(S) - x(S)$ arranged according to magnitude, where S runs over all coalitions of N; i.e., $i < j \rightarrow \theta^i(x) \geq \theta^j(x)$. The θ vectors are ordered lexicographically as follows: if $\theta(x)$ precedes $\theta(y)$ in the ordering, then either $\theta^1(x) < \theta^1(y)$, or $\theta^2(x) < \theta^2(y)$ and $\theta^1(x) = \theta^1(y)$, or $\theta^3(x) < \theta^3(y)$ and $\theta^j(x) = \theta^j(y)$ for $j = 1$ and $2, \ldots$, or $\theta^{2^n}(x) < \theta^{2^n}(y)$ and $\theta^j(x) = \theta^j(y)$ for $j = 1, \ldots, 2^n - 1$. This lexicographic order on R^{2^n} induces a quasi order on R^n, and the nucleolus is the first payoff vector in the quasi order. If the scalar $v(S) - x(S)$ is considered the objection raised by coalition S against payoff x, then the nucleolus is that payoff where the maximum objection raised over all coalitions is less than the maximum objection raised over all coalitions against any other payoff. It should be clear that the nucleolus is unique and in the core if the core is nonempty.

The nucleolus can be calculated using linear programming.[8] For our example, the nucleolus is $x_1 = 133.8$, $x_2 = 120.0$, and $x_3 = 46.2$, and is represented by point n^* in Figure 1. To further appreciate the rationale behind the nucleolus, it can be compared to the Shapley value payoff in the same example. Let s^* and n^* be the Shapley value and nucleolus payoff vectors (i.e., $s^* = (s_1^*, s_2^*, s_3^*) = (144.6, 114.6, 40.8)$ and $n^* = (n_1^*, n_2^*, n_3^*) = (133.8, 120.0, 46.2)$ so that the θ vectors of objections are as shown in Table 1.[9]

Coalitions $\{1\}$, $\{1,2\}$, and $\{1,3\}$ favor the Shapley value, and coalition $\{1,2\}$ has the greatest objection (−46.2) to the nucleolus. On the other hand, coalitions $\{3\}$, $\{2\}$, and $\{2,3\}$ favor the nucleolus and coalition $\{3\}$ has the greatest objection (−40.8) to the Shapley value. The nucleolus is superior in the sense that the objection of −46.2 against the nucleolus is weaker than the objection of −40.8 against the Shapley value.

As it stands, the objection raised by

TABLE 1—A COMPARISON OF THE SHAPLEY VALUE AND THE NUCLEOLUS

S	$v(S)$	$n^*(S)$	$s^*(S)$	$\theta(n^*)$	$\theta(s^*)$
N	300 0	300.0	300.0	0	0
{3}	0	46.2	40.8	−46.2	−40.8
{1, 2}	207.7	253.8	259.2	−46.2	−51.5
{2}	0	120.0	114.6	−120.0	−114.6
{1, 3}	60 0	180 0	185.4	−120.0	−125.4
{1}	0	133.8	144.6	−133.8	−144.6
{2, 3}	0	166.2	155.4	−166.2	−155.4

coalitions against certain payoffs are independent of the number of players in the coalitions. A reasonable alternative is to allow objections by large coalitions to carry more weight than objections by small coalitions. This may be especially important in light of the asymmetric role of the entrepreneur. His objection could be either amplified or discounted depending on his bargaining power.

Both the Shapley value and nucleolus are guaranteed to be in the core of the pricing game, and they both yield lump sum charges that can be reinterpreted as two-part tariffs. Price is equal to marginal cost since the characteristic function implies that each consumer receives $q_i(N)$. The set of license fees, however, depends on the shape of the demand and cost curves. The core can be used to indicate the bounds on these license fees. Using (3), (4), and $g(i) = pq_i(N) + T_i$, we have for $S \subseteq N$,

$$\sum_S T_i \leq C(q(S)) - \sum_S pq_i(N) + \sum_S [A_i(q_i(N)) - A_i(q_i(S))]$$

Thus, the set of upper bounds indicate that the license fees for subcoalition S must not be so great as to result in a net benefit to S, less than what S could acquire on its own. For a lower bound, the standard nonnegativity constraint on T_i continues to hold. To see this, consider the outcome if $T_j < 0$, i.e., the charge to consumer j is less than $pq_j(N)$. Then

$$\sum_{N-\{j\}} g(i) > C(q(N)) - pq_j(N)$$

[7] In case $v(i) > 0$ for any $i \in N$, the game must be normalized so that the above conditions imply individual and group rationality.

[8] See Owen for this linear programming method.

[9] Slight discrepancies are due to rounding.

FIGURE 2

and $\sum_{n-\{j\}} A_i(q_i(N)) - C(q(N))$

$$+ pq_j(N) > \sum_{N-\{j\}} x_i$$

With (4), the condition for x to be in the core, this implies

$$(10) \quad \sum_{N-j} [A_i(q_i(N)) - A_i(q_i(N - \{j\}))] > C(q(N)) - C(q(N - \{j\})) + pq_j(N)$$

However, this last inequality is false as we illustrate diagrammatically in Figure 2. Substituting the lettered areas into (10) yields

$$A + B + C - A - B > B + C + D + E + F - B - E - F$$

But $C \not> C + D$, implying that each consumer must at least pay the uniform price for the quantity he consumes.

IV. Concluding Remarks

In this paper, we have examined the familiar problem of pricing in decreasing cost industries. In doing so, we have emphasized a game-theoretic criterion for evaluating traditional pricing arrangements. This criterion is the stability concept which must be satisfied if consumers and producers are to accept charges as levied, and not disband into smaller groups. It was seen that game theory provided a natural environment for analyzing these traditional pricing arrangments, and the pricing game developed was shown to be convex. Finally, it was shown that well-known pricing arrangements such as two-part tariffs do not guarantee stability, while certain game-theoretic solutions, including the Shapley value and nucleolus, do guarantee stability.

APPENDIX

We first prove that the characteristic function for a pricing game (PG) without the entrepreneur satisfies (2). In PG, indices $1, \ldots, n$ refer to consumers and consumer coalitions can produce the good. Thus (1) becomes

$$v(S) \max_q \{A_S(q) - C(q), q \geq 0\}$$

for all $S \subseteq N$ and $q = \Sigma_S q_i$. The proof is divided into two cases. Case 1 considers $q(S \cup k) \leq q(T)$, and Case 2 considers $q(S \cup k) > q(T)$.[10]

CASE 1: $\quad q(S \cup k) \leq q(T)$

In what follows, we will make use of the following observations. If a consumer moves from his present coalition to a new coalition, where the new coalition consumes more in total than the present one, then the consumer's own consumption will never decrease. Thus $i \in S \subset T$ implies $q_i(T) \geq q_i(S)$. To see this, note that for $q(T) \geq q(S)$, we have

$$c(q(S)) \geq c(q(T)) \text{ or } \Psi_S(q(S)) \geq \Psi_T(q(T))$$

Since demand functions are monotonically decreasing we have

$$q_i(S) = \Psi_i^{-1}(\Psi_S(q(S))) \leq \Psi_i^{-1}(\Psi_S(q(T))) = q_i(T)$$

If each consumer $i \in S$ increases his quantity demanded from $q_i(S)$ to $q_i(S \cup k)$, the gain in benefits is

[10] For convenience, the notation "$S \cup k$" and "$T \cup k$" is used throughout the Appendix instead of the correct notation "$S \cup \{k\}$" and "$T \cup \{k\}$."

$$\sum_S [A_i(S \cup k) - A_i(q_i(S))]$$

The increased cost of such a move is

$$C\left(\sum_S q_i(S \cup k)\right) - C(q(S)) = C(q(S \cup k) - q_k(S \cup k)) - C(q(S))$$

which must be at least as great as the increased benefits, because $q_i(S)$ was chosen to maximize welfare. Therefore,

$$\sum_S [A_i(q_i(S \cup k)) - A_i(q(S))] \leq C(q(S \cup k) - q_k(S \cup k)) - C(q(S))$$

and adding $A_k(q_k(S \cup k)) - C(q(S \cup k))$ to both sides then rearranging yields

$$v(S \cup k) - v(S) \leq A_k(q_k(S \cup k)) - C(q(S \cup k)) - C(q(S \cup k) - q_k(S \cup k)) = B_1$$

Next, consider the coalition $T \cup k$. If we subtract a nonnegative quantity from each consumer $i \in [T \cup k]$, the benefits lost must exceed the decrease in cost; otherwise the coalition is not maximizing $v(T \cup k)$. Thus

$$\sum_{T \cup k} A_i(q_i(T \cup k)) - \sum_T A_i(q_i(T)) - A_k(q_k(S \cup k)) \geq [C(q(T \cup k)) - C(q(T) + q_k(S \cup k))]$$

and adding $C(q(T))$ to both sides and rearranging yields

$$v(T \cup k) - v(T) \geq A_k(q_k(S \cup k)) - [C(q(T) + q_k(S \cup k) - C(q(T))] = B_2$$

We need only show that $B_2 \geq B_1$ or

$$C(q(S \cup k)) - C(q(S \cup k) - q_k(S \cup k)) \geq C(q(T) + q_k(S \cup k)) - C(q(T))$$

But this follows from the assumption that c is nonincreasing and $q(S \cup k) \leq q(T)$. Therefore, (2) holds for $q(S \cup k) \leq q(T)$.

CASE 2: $q(S \cup k) > q(T)$

Using an argument from Case 1 where consumers in S increase quantity demanded beyond $q_i(S)$, we have

$$\sum_S [A_i(q_i(T)) - A_i(q_i(S))] \leq \left[C\left(\sum_S q_i(T)\right) - C(q(S))\right]$$

Adding

$$B_3 = \sum_S A_i(q_i(S \cup k)) + A_k(q_k(S \cup k)) - \sum_S A_i(q_i(T))$$

and $-C(q(S \cup k))$ to both sides yields

$$v(S \cup k) - v(S) \leq B_3 - C(q(S \cup k)) + C\left(\sum_S q_i(T)\right) = B_4$$

Again using an argument from Case 1 where a nonnegative quantity is subtracted from each consumer $i \in [T \cup k]$, we obtain

$$\sum_S [A_i(q_i(T \cup k)) - A_i(q_i(S \cup k))] + A_k(q_k(T \cup k)) - A_k(q_k(S \cup k)) + \sum_{T-S} [A_i(T \cup k)) - A_i(q_i(T))] \geq C(q(T \cup k)) - C\left(q(S \cup k) - \sum_S q_i(T) + q(T)\right)$$

Adding $C(q(T)) - \Sigma_S A_i(q_i(T))$ to both sides and rearranging yields

$$v(T \cup k) - v(T) \geq B_3 - C\left(q(S \cup k) + q(T) - \sum_S q_i(T)\right) + C(q(T)) = B_5$$

We need only show that $B_5 \geq B_4$, or

$$C(q(T)) - C\left(\sum_S q_i(T)\right) \geq C\left(q(S \cup k) + \sum_{T-S} q_i(T)\right) - C(q(S \cup k))$$

But this follows from the assumption that c is nonincreasing and $q(S \cup k) > q(T)$. Therefore, (2) holds for $q(S \cup k) > q(T)$.

This completes the proof that PG is convex. Although we assumed continuity on c

and the demand functions, this can be relaxed to piecewise continuity. Finally, if the entrepreneur is included, the characteristic function is still convex. We state this in the following proposition, the proof of which is not difficult.

PROPOSITION 2: *If v is a convex game on N and w is a game on $N \cup \{j\}$ given by $w(S) = 0$, $w(S \cup j) = v(S)$, $w(j) = 0$ for $S \subseteq N$, then w is convex.*

REFERENCES

J. M. Buchanan, "An Economic Theory of Clubs," *Economica*, Feb. 1965, *32*, 1–14.

——, "Peak Loads and Efficient Pricing: Comment," *Quart. J. Econ.*, Aug. 1966, *80*, 463–71.

D. Gatley, "Sharing the Gains from Regional Cooperation: A Game Theoretic Application to Planning Investment in Electric Power," *Int. Econ. Rev.*, Feb. 1974, *15*, 195–208.

S. C. Littlechild, "A Game-Theoretic Approach to Public Utility Pricing," *Western Econ. J.*, June 1970, *2*, 162–66.

—— **and G. Owen,** "A Simple Expression for the Shapley Value in a Special Case," *Manage. Sci.*, Nov. 1973, *20*, 370–72.

E. T. Lochman and A. B. Whinston, "A New Theory of Pricing and Decision Making in Public Investments," *Bell J. Econ.*, Autumn 1971, *2*, 606–28.

R. Duncan Luce and Howard Raiffa, *Games and Decisions*, New York 1957.

W. Y. Oi, "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quart. J. Econ.*, Feb. 1971, *85*, 77–96.

G. Owen, "A Note on the Nucleolus," *Int. J. Game Theory*, 1974, *3*, 101–03.

D. Patinkin, "Demand Curves and Consumer's Surplus," in Carl Christ et. al., eds., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford 1963.

M. V. Pauly, "Clubs, Commonality, and the Core: An Integration of Game Theory and the Theory of Public Goods," *Economica*, Aug. 1967, *34*, 314–24.

John Rawls, *A Theory of Justice*, Cambridge, Mass. 1971.

N. Ruggles, "Recent Developments in the Theory of Marginal Cost Pricing," *Rev. Econ Studies*, No. 2, 1950, *17*, 107–26.

D. Schmeidler, "The Nucleolus of a Characteristic Function Game," *SIAM J. Appl Math.*, Nov. 1969, *17*, 1163–70.

L. S. Shapley, "The Value of an N-Person Game," in Harold W. Kuhn and A. W. Tucker, *Contributions to the Theory of Games*, Princeton 1953.

——, "Cores of Convex Games," *Int. J. Game Theory*, 1971, *1*, 11–26.

M. Shubik, "Incentives, Decentralized Control, the Assignment of Joint Costs and Internal Pricing," *Manage. Sci.*, Apr. 1962, *8*, 325–43.

Gold, Dollars, Euro-Dollars, and the World Money Stock under Fixed Exchange Rates

By ALEXANDER K. SWOBODA*

In a rapidly inflating domestic economy, explanation of the inflationary process must of necessity focus on the determinants of the money supply. By analogy, in a closely integrated world economy under fixed exchange rates the behavior of the sum of individual countries' money stocks—the "world money stock"—should play an important role in determining the behavior of the "world price level" (an index of national price levels). This is recognized in analytical models of the international monetarist variety where, under the assumption that all goods are traded (or more generally that relative prices are not affected in the long run by monetary disturbances), the world price level adjusts to equate the world demand for money with the supply.

Strictly fixed *exchange rates* imply that national money stocks can be treated as components of a Hicksian composite commodity, the world money stock, since exchange-stabilization operations prevent variations in the relative values of national currencies. Closely integrated *capital markets* insure that the world money stock is redistributed rapidly from country to country in response to payments disequilibria of monetary origin, thus ensuring a tendency towards rapid return to *balance-of-payments equilibrium* (which can, of course be frustrated by systematic attempts at neutralization of reserve flows). Closely integrated *goods markets* insure that the *price levels* of various countries move in harmony—abstracting of course, from divergent trends in productivity and/or tastes that may cause changes in relative prices, including both the terms of trade and the ratio of the price of nontraded to traded goods. In such a world, one can view the world stock of money as determining the world price of a composite commodity, the components of which are national output levels. Though far too simple for many purposes, this Humean or Ricardian view of the world economy is instructive in periods dominated by disturbances of monetary origin.

For this type of analysis to be complete, however, the question of what determines the supply of money in the world must be answered. This paper seeks to answer this question within the confines of a conceptually (though not necessarily algebraically) very simple model. The world is assumed to be divided into two parts, Europe and the United States. National money stocks consist of commercial bank liabilities only, and the world money stock is defined as the sum of the money balances held by the public of each country.[1] Various institutional arrangements are considered, including a gold standard, a dollar standard, and the Euro-dollar system. The model provides a first answer to such questions as: does it make any difference to inflation whether monetary expansion originates in one region or the other; what asymmetries does a dollar standard introduce into the international monetary system; what determines the size of the Euro-dollar market and in what sense, if any, is its growth inflationary?

Two key assumptions are used to answer these questions, namely, 1) that reserve

*Professor of international economics, Graduate Institute of International Studies, Geneva, and visiting professor of economics, Harvard University. The original version of this paper was prepared for the Money Study Group's Oxford Seminar in honor of James Meade, September 25–27, 1974. It includes material developed in connection with a research project on "National Economic Policy and the International Monetary System" at the Graduate Institute of International Studies under a grant from the Ford Foundation.

[1] To sum national money stocks, they must of course be expressed in terms of the same currency, existing fixed exchange rates providing the required conversion factor.

flows take place until balance-of-payments equilibrium is reestablished and 2) that payments equilibrium requires that the money stocks of the two regions stand in given proportion to each other. These assumptions are not arbitrary. They underlie the Hume-Senior-Ricardo demonstration that the natural distribution of specie (here the natural distribution of the world money stock) tends to assert itself through the specie flow (here the reserve flow) mechanism. With these assumptions it is possible to analyze the world money supply process under fixed exchange rates in a manner that is strictly analogous to the analysis of closed-economy monetary systems characterized by various stable asset-preference ratios. The model thus integrates, in simplified fashion to be sure, the money supply analysis pioneered by James Meade (1934) with balance-of-payments theory to which he has contributed so much.[2]

It should be evident that the two assumptions above are particularly useful when analyzing a world of strictly fixed exchange rates where spot rates are not expected to change and are equal to forward rates. The present paper is confined to this case on several grounds. In the first place, this is not a bad assumption to make when interpreting, in broad and global terms, the evolution of the international monetary system until the breakdown of the "Bretton Woods system." Second, the analysis developed below should help throw light on the issues involved in designing a fixed exchange rate system, should the world return to such a system; furthermore, it remains relevant to the extent that managed floating retains elements of fixed rates and that holdings of foreign-currency assets by the public and authorities introduce elements of interdependence among national banking systems even under formally floating exchange rates.[3] Be that as it may, the virtue of the procedure adopted in this paper is that it makes possible the analysis of problems that have not been solved satisfactorily hitherto. For instance, the determinants of the base of Euro-dollar expansion are usually assumed to consist of exogenously determined flows of deposits to that market, these flows depending in turn, in the more sophisticated existing analyses, on exogenously given payments disequilibria. In the present paper these flows are made endogenous. The only exogenous variables are the quantities of domestic assets of central banks (these are policy variables) and the world stock of outside assets (gold).

The analysis begins with the discussion of a simple gold standard model and of a number of key assumptions. A more general model is then presented; its complexity motivates the procedure adopted in subsequent sections of discussing specialized versions of the basic general model. These versions include the dollar standard and introduce the Euro-dollar market. The role of neutralization operations is emphasized throughout and the importance of institutional arrangements for the effectiveness of monetary policy is illustrated with a number of numerical examples in Section V. For instance, on not entirely unlikely assumptions as to the magnitudes of behavior parameters, a $1 open-market purchase of securities in the United States increases the world money stock by some $5.9, whereas an equivalent purchase originating in the rest of the world increases the world money stock by only some $3.5. If the United States neutralizes all reserve flows, these figures are changed to 10 and 0, respectively. The concluding section of the paper discusses a number of implica-

[2]Money supply analysis of the multiplier variety has since become a standard feature of textbooks in money and banking; for some of the main contributions to its development see, among others, Philip Cagan, Milton Friedman and Anna Schwartz, and Karl Brunner and Alan Meltzer. The *locus classicus* of Meade's contribution to balance-of-payments theory is of course his *The Balance of Payments*.

[3]Analysis of such interdependence can, however, become quite complex. The determinants of the demand for foreign-currency holdings are not easy to specify and much depends on the specific assumptions made as to the *modus operandi* of foreign-exchange intervention by central banks. There are, however, some obvious extensions of the analysis of Section IV below; these are left to the reader.

tions of the analysis.[4] For convenience, a list of symbols and of the main behavioral relationships and balance sheet identities is provided in the Appendix.

I. The World Money Stock under a Gold Standard

The world is assumed to consist of two countries, the United States (country 1) and Europe (country 2). Variables pertaining to Europe are identified by an asterisk. The exchange rate between the two currencies is assumed to be strictly fixed (both spot and forward), without margins, and is assumed for convenience to be equal to 1. This allows us to reckon all amounts in dollars.[5] The world money stock is defined as the sum of the money supplies *in the hands of the public* resident in each country.[6] Commercial bank liabilities are the only type of money held by the public. No distinction is made between demand and time deposits. Commercial banks hold a fixed ratio of reserves against their deposit liabilities. We thus neglect the effect of a change in money supplies on interest rates and the feedback to desired reserve ratios; this is not a serious defect from our point of view since the qualitative results hold under variable interest rates as long as the latter are stable functions of money supplies and as long as reserve ratios are stable functions of interest rates.

[4]I have made use of a similar general analytic technique in a previous paper. That paper, however, suppresses the algebra made explicit in the present one and addresses itself to a separate though related set of questions In a recent paper, Hans Genberg and I have estimated a simple, simultaneous two-region model of worldwide inflation under the dollar standard for the period 1957 71 This model which contains a rudimentary world money supply process confirms the asymmetries noted in the present paper, and offers some evidence as to the process of adjustment from short to long run

[5]For some purposes, notably the analysis of devaluation, the exchange rate can easily be introduced in the formulae developed below

[6]Holdings of foreign-currency denominated money by the public are ignored in this section. One special form of such holdings, Euro-dollar deposits of the European public, is discussed in subsequent sections

The method of analysis is that of comparative statics. An initial equilibrium is disturbed by a change in an exogenous variable or by an autonomous shift in a behavior parameter, and the new equilibrium is compared with the initial one. There are three exogenous variables in the system—two policy variables, namely, the domestic assets held by the United States (A) and the European (A^*) central banks and a "nature-given" variable, the world stock of gold (G). Behavior parameters reflect desired or compulsory reserve ratios of commercial banks, and asset preferences of the public. An increase in A represents an expansionary monetary policy by the U.S central bank, that is, an open-market purchase of securities: an increase in A^* represents a similar policy by the European central bank. Endogenous variables include the world money stock, its distribution among the residents of the two countries, and the distribution of foreign-exchange reserves among the two central banks.

Equilibrium is defined by equality of the demand and supply of money in each country and by payments equilibrium. In the simple static models of this paper equilibrium obtains when the asset preferences of the public are satisfied and the world money stock has been distributed among the residents of the two countries in the proportion required for payments equilibrium. This proportion is assumed to be fixed and is denoted by β (to be discussed in the next paragraphs). Endogeneity of foreign-exchange reserves makes attainment of this equilibrium proportion possible; the implicit dynamic mechanism of adjustment is that reserves flow until the distribution of the world money stock compatible with payments equilibrium has been reached.

As the existence of such an equilibrium distribution of the world money stock is crucial to the analysis, a brief discussion of its meaning is in order here: β is defined as the ratio of the first country's money stock to the world money stock. In equilibrium the demand for money is equal to the supply in both countries. Algebraically, equations (1), (2), and (3) below define, respectively, the

equality of the demand and supply of money in countries 1 and 2, and the equilibrium distribution β of the world money stock:

$$(1) \quad M = L(Y,i)\cdot P$$

$$(2) \quad M^* = L^*(Y^*,i^*)\cdot P^*$$

$$(3) \quad \beta = \frac{M}{M_w} = \frac{M}{M + M^*} = \frac{L(Y,i)\cdot P}{L(Y,i)\cdot P + L^*(Y^*,i^*)\cdot P^*}$$

where M is the money stock, L is the demand for money, Y is output, i is the rate of interest, and P the price level. The subscript w identifies world variables

It is obvious that in the equilibrium of a classical static world where Y, Y^*, i, i^* are given and where P and P^* are equalized through trade (or change in proportion in the absence of long-run changes in relative prices) β can take on one and only one value (the world price level adjusts to changes in the world money stock). More generally, however, it can be shown that β will, for a large class of models, be invariant to the origin of money supply disturbances and that it will either be invariant with respect to a change in M_w or be systematically and predictably related to such a change. Differentiating (3) and denoting percentage changes in a variable by capping it with a hat, one obtains:

$$(4) \quad \hat{\beta} = (1-\beta)\{(\eta_{LY}\hat{Y} - \eta^*_{L^*Y^*}\hat{Y}^*) + (\eta_{Li}\hat{i} - \eta^*_{L^*i^*}\hat{i}^*) + (\hat{P} - \hat{P}^*)\}$$

where η_{yx} denotes the elasticity of y with respect to x. In a fixed price two-country model of the Keynesian variety, it can be shown that β will still be independent of the national origin of a world money stock change.[7] In addition $\hat{Y} = \hat{Y}^*$ if the income elasticity of the demand for imports is unity in both countries and payments equilibrium requires that $\hat{i} = \hat{i}^*$ (if capital flows are a function of the ratio of interest rates). In these circumstances, β does not change in response to a change in the world money stock if the income and interest elasticities of the demand for money are the same across countries. Otherwise, an expansion of the world money stock will raise β if $\eta_{LY} > \eta^*_{L^*Y^*}$ and if $\eta_{Li} < \eta^*_{L^*i^*}$.

In what follows, it will be assumed for simplicity that, in equilibrium, β can be treated as exogenous, its long-term changes being governed by such factors as different rates of growth of income and productivity across countries and sectors. Such changes in β due to changes of nonmonetary origin occur continuously in the real world but will be abstracted from in this paper.[8] Of course, β will only be established in the long run, that is, after complete payments adjustment has taken place. At any point in time, the actual value of β is subject to short-run changes due to variations in monetary policy. It tends back to its initial equilibrium through a complex adjustment mechanism which involves reserve flows, changes in the world and national money stock, and variations in interest rates, prices, and income levels. The relevance of the analysis depends partly on the speed of that adjustment mechanism—which is likely to be quite rapid in a world of integrated goods and capital markets.[9]

With these assumptions, analysis of the determinants of the world money stock is straightforward. Consider, first, a version of the gold standard. The only international

[7]This is demonstrated, and is shown to hold independently of the degree of capital mobility, in the author and Dornbusch

[8]This need not be a cause of great concern here since 1) the purpose of the analysis is to trace the consequences of various institutional arrangements for disturbances of monetary origin, 2) adjustment in an integrated world economy tends to be fairly rapid in chronological time as noted in the introduction, 3) the consequences of a change in β due to changes in relative prices or differential growth rates can be readily traced, 4) the analysis is designed, in part, to illuminate adjustment in times dominated by disturbances originating on the supply side of the monetary process

[9]Empirical evidence on the speed of adjustment is provided in Genberg and the author The mean time lag of the adjustment of β to its equilibrium value is estimated to be two quarters in a simple simultaneous two-region model of the industrialized world for the period 1957 71, which incorporates a rudimentary world money supply process.

reserve asset held by central banks is gold, the world stock of which is assumed to be given exogenously. In the particular version of the gold standard that follows, it is assumed that central banks do not keep a fixed ratio of gold to the national money stock. Instead, they exercise a certain amount of policy independence by engaging in open-market purchases and sales of domestic assets.[10] The structure of the system is given in the balance sheets below.

United States

Fed		Commercial Banks	
Assets	Liabilities	Assets	Liabilities
A	R	R	M
αG		L_I	

Europe

European Central Bank		Commercial Banks	
Assets	Liabilities	Assets	Liabilities
A^*	R^*	R^*	M^*
$(1-\alpha)G$		L_I^*	

In addition to the symbols already defined above, R represents commercial bank reserves with their central bank, L_I loans and investments of commercial banks, and α the proportion of the world's gold stock G, held by the $U S$ central bank (the Fed). Note that α is an *endogenous* variable. Three further definitions are required:

$$r = \frac{R}{M} = \frac{1}{m} \tag{5}$$

$$r^* = \frac{R^*}{M^*} = \frac{1}{m^*} \tag{6}$$

$$\beta = \frac{M}{M_w} = \frac{M}{M + M^*} = \frac{m(A + \alpha G)}{mA + m^*A^* + m^*G + (m - m^*)\alpha G} \tag{7}$$

[10]That independence is, obviously, limited by the requirement that gold holdings be positive Other versions of the gold standard can readily be worked out For instance, one could easily develop a model in which national money supplies are, in the long run, proportional to central banks' holdings of gold and where central banks' portfolios of domestic assets are varied according to some stock adjustment rule towards that proportion in the short run

Equations (5) and (6) define the reserve ratios of the two commercial banking systems (assumed to be fixed and equal by definition to the inverse of the national money supply multipliers m and m^*). Equation (7) restates the definition of β, the equilibrium distribution of the world money stock.

Suppose that the Fed carries out an expansionary open-market operation, increasing A by dA. The *impact* effect is to increase M, and hence M_w, by mdA. This is not the end of the story, however, since, other things equal, the open-market operation has created an excess supply of money in the United States. That country experiences a payments deficit; as a consequence gold flows to Europe, increasing the money supply there and decreasing it in the United States. To obtain the *final* effect on the world money stock and other variables after reserve flows have restored payments equilibrium, it is convenient to solve for M_w in terms of exogenous variables and behavior parameters to yield:[11]

$$M_w = \frac{A + A^* + G}{r\beta + r^*(1 - \beta)} \tag{8}$$

A number of important conclusions are immediately apparent from this expression. First, and most important, the effect on the world money supply of an equal size increase in A, A^*, or G is exactly the same: the final effect on M_w of an increase in "base money" is independent of its national origin. A "world money base" $(A + A^* + G)$ can meaningfully be defined. There is a basic symmetry in a gold standard system which insures that an open-market operation has the same effect on the world money stock—and, hence, on the world level of economic activity or prices—wherever it originates. Redistribution of gold through the payments adjustment mechanism insures that monetary policy becomes internationalized in symmetric fashion. Second, the "world money multiplier" (1 over the denominator on the right-hand side of equation (8)) is a weighted

[11]The derivation of expression (8) is given in fn 26.

TABLE 1

	General Case	Equal Reserve Ratios $\left(r = r^* = \frac{1}{m} = \frac{1}{m^*}\right)$	Neutralization by Country 1	
$\frac{dM_w}{dA} = \frac{dM_w}{dA^*}$	$\frac{1}{r\beta + r^*(1-\beta)}$	m	$\frac{dM_w}{d\bar{B}} = \frac{m}{\beta}$	$\frac{dM_w}{dA^*} = 0$
$\frac{dM}{dA} = \frac{dM}{dA^*}$	$\frac{\beta}{r\beta + r^*(1-\beta)}$	$m\beta$	$\frac{dM}{d\bar{B}} = m$	$\frac{dM}{dA^*} = 0$
$\frac{dM^*}{dA} = \frac{dM^*}{dA^*}$	$\frac{(1-\beta)}{r\beta + r^*(1-\beta)}$	$m(1-\beta)$	$\frac{dM^*}{d\bar{B}} = \frac{(1-\beta)}{\beta} m$	$\frac{dM^*}{dA^*} = 0$
$\frac{dIR^*}{dA}$	$\frac{r^*(1-\beta)}{r\beta + r^*(1-\beta)}$	$(1-\beta)$	$\frac{dIR^*}{d\bar{B}} = \frac{(1-\beta)}{\beta}$	
$\frac{dIR^*}{dA^*}$	$\frac{-r\beta}{r\beta + r^*(1-\beta)}$	$-\beta$		$\frac{dIR^*}{dA^*} = -1$

average of national money multipliers, the weights being the relative economic sizes of the two countries, β and $(1 - \beta)$. This result appeals to common sense. Suppose the United States to be very large relative to Europe: β tends towards 1. Any open-market operation will tend to change mainly the money supply of the United States and hence its multiplier should dominate.

The symmetry noted above as well as the role of size is brought out very clearly in Table 1, which lists the effects of a change in A and A^* on M, M^*, M_w, and on IR^*, the stock of international reserves held by country 2, here gold. The first column of the table gives results for the general case, the second for the special case where the money multipliers are the same in the two countries and the common multiplier and reserve ratio are denoted by m and r, respectively (ignore the third column for the moment).

Equality of dM_w/dA and dM_w/dA^* is the one conclusion from which all other results in Table 1 follow. This equality obtains even if domestic money multipliers differ. Suppose, for instance, that $m^* > m$. A European open-market purchase of securities dA^* increases the world money stock by more than a corresponding purchase in the United States dA, *before* reserves flow (at impact). But, when A^* increases, Europe experiences a payments deficit that reduces Europe's money supply by more than it increases America's, and vice versa when A increases. This is why $dM_w/dA = dM_w/dA^*$ even if $m \neq m^*$. It immediately follows that the *final* effect of an open-market operation on individual national money stocks (M and M^*) is independent of its origin; for, the given change in the world money stock is distributed among countries 1 and 2 in the fixed proportions β and $1 - \beta$, respectively, that is, in proportion to their relative economic size. Reserve changes are also proportional to size, but depend in addition on money multipliers when the latter differ. The higher r (given r^*) and the lower r^* (given r), the smaller will be the U.S. reserve loss attendant on an American open-market purchase of securities of given size: a high r reduces the initial excess supply of money created by dA and a lower r^* implies that a small redistribution of *reserves* towards Europe suffices to produce a large increase in the European *money supply*. For similar reasons, the higher r^* (given r) and the lower r (given r^*), the smaller will be the European reserve loss attendant upon expansionary European monetary policy.

The role of size is highlighted by considering the case where money multipliers are

equal. Open-market purchases of securities increase the domestic money supply in proportion to the country's relative economic size. They cause reserve losses that are inversely proportional to the country's relative economic size. Consider the special case where Europe becomes negligibly small relative to the United States, that is, β tends towards unity. A European open-market purchase (sale) results in an equal loss (gain) of foreign-exchange reserves. As a corollary, a European open-market operation fails to affect the money supply of a sufficiently small Europe. The explanation is simple: an open-market operation becomes generalized and serves to increase the world money supply the small country only retaining (or receiving) its infinitesimally small share of the total change. Our model thus demonstrates, in slightly different guise, the standard conclusion of analyses of the small open economy, namely, that, except in the short run, the monetary authorities of such economies have no control over the national money stock, an open-market sale or purchase resulting in a countervailing reserve loss. How short the short run is, is, of course, of crucial importance for the conduct of monetary policy For a very small open economy, in a world of closely integrated goods and capital markets, it is likely to be quite short in chronological time.[12]

Neutralization operations by the monetary authorities in one country effectively reduce the role of monetary policy in the other to what it would be were that other country infinitesimally small. Suppose that the United States buys (sells) an equivalent amount of securities in the open market whenever it loses (gains) gold. Consider the effect of an open-market purchase of securities by the European central bank under this assumption. The increase in A^* results initially in an increase in M^* and an outflow of gold to the United States. American authorities, however, prevent this reserve gain from affecting the *U.S.* money supply by decreasing A in step. A new equilibrium is reached when A has decreased by the same amount as A^* initially increased, and Europe has lost an equivalent amount of gold to the United States. Formally, neutralization can be modelled by letting $\overline{B} = A + \alpha G$, where $\overline{B}$ is the level at which the *U.S.* monetary authorities maintain the monetary base. This implies that A becomes an endogenous variable and that $dA = -d\alpha G$. The reserves of the *U.S.* commercial banking system are kept equal to $\overline{B}$, and the world money supply formula (8), for the case where $m = m^*$ becomes:

$$(8') \qquad M_w = \frac{m}{\beta} \cdot \overline{B}$$

In equilibrium, the world money stock is determined entirely by the level at which the neutralizing authorities choose to keep their money stock ($m\overline{B}$) together with the latter's share in the world money supply, that is, by the neutralizing country's relative economic size. The last column of Table 1 lists results for the case where the United States neutralizes reserve flows and confirms the conclusion that neutralization by the United States confines European open-market operations to effecting offsetting reserve flows and robs them of any effect on the European money stock.[13] In contrast, the *U.S.* monetary policy becomes quite powerful. An increase in the *autonomous* component $\overline{B}$ of domestic assets held by the Fed raises the world money stock by the American domestic money multiplier

[12] Both analytical and empirical reasons underlie the statement in the text For empirical evidence on high though not unitary offset coefficients, see, for instance, Pentti Kouri and Michael Porter In Genberg and the author, it is estimated that it took an average of two years during the 1957 71 period for monetary policy in the non-*U.S.* industrialized world as a whole to be completely offset by flows of international reserves when the United States pursued a policy of neutralizing foreign influences on the American monetary base. This is not to deny that even a small country can maintain for some time a money stock that implies a payments disequilibrium. The required neutralization operations, however, are likely to become extremely large and unsustainable in practice in a world of integrated goods and capital markets, if the policy is pursued systematically over time unless it is designed to smooth out short-run variations in the demand and supply of money that tend to cancel out over time (the cycle)

[13] The derivatives in the table have to be reinterpreted as being taken with respect to $d\overline{B}$ and not dA in the case of neutralization.

times $1/\beta$. The total increase in A is of course larger since the United States neutralizes the reserve loss attendant on an increase in $\bar{B}$.[14] The smaller the United States is relative to Europe, the larger will be the reserve loss, and hence, the required neutralization operation. The latter tends to infinity as the neutralizing country becomes negligibly small relative to the rest of the world. This is in accord with yet another standard result of theorizing about small open economies, namely, that neutralization becomes impossible when the mobility of capital is perfect. Here, however, it is not the *rate* of neutralization operations per unit of time but the absolute size of the neutralization operation compatible with full equilibrium that becomes infinite and this conclusion is established independently of the degree of capital mobility.[15]

The ability of countries to increase their money stock and to neutralize the resulting payments deficit does depend of course on the size of their stock of international reserves, and, under the gold standard, that stock is limited. Furthermore, both countries can play the neutralization game. If they do, monetary equilibrium in the world economy cannot obtain, except perchance. The return to equilibrium that would be brought about by the effect of redistribution of reserves on national money supplies after any monetary disturbance is resisted by both countries. Neither country is willing to accept the burden of international adjustment, payments imbalances become self-perpetuating, and the world economy enters what Robert Mundell has called "the international disequilibrium system."

II. A More General Model of World Money Stock Determination

The gold standard model set out above is a very special case of the more general one presented in this section. The gold standard model was analyzed in some detail since its very simplicity reveals clearly some of the features of world money stock analysis that underlie results derived from the more general model but whose essential movitation is hidden by the latter's greater complexity. A detailed description of the algebraic relationships that define the more general model is relegated to the Appendix as its structure can be grasped intuitively from a study of the balance sheet pattern it assumes and from a few words of explanation.

United States

Fed		Commercial Banks	
Assets	Liabilities	Assets	Liabilities
A	R	R	M'
αG	D^*	L_I	D_2^*
			DEB

Europe

European Central Bank		Commercial Banks	
Assets	Liabilities	Assets	Liabilities
A^*	R^*	R^*	M'^*
$(1-\alpha)G$		L_I^*	ED^*
D^*		DEB	ED
D_2^*			CED
CED			

This model differs from the gold standard one principally by recognizing that Europe's central bank can hold a variety of dollar assets as foreign-exchange reserves in addition to gold and by allowing for the existence of the Euro-dollar market. As six new types of asset holdings are introduced, six new behavior relationships (or equilibrium conditions) must be added to the preceding ones in order to obtain a determinate equilibrium value of the world money stock and of its distribution.

The European central bank can hold its foreign-exchange reserves in four forms: gold, $(1-\alpha)G$, as before; dollar deposits with the Fed, D^*; dollar deposits with *U.S.* commercial banks, D_2^*; dollar deposits in the Euro-dollar market, that is, with European commercial banks, *CED*. It is assumed that the European central bank

[14] As a matter of fact, $dA/d\bar{B} = 1/\beta$

[15] Note, however, that the time required for the equilibrium distribution of the world money stock to assert itself does depend on the degree of capital mobility.

first decides which proportion of its total foreign-exchange reserves to keep in gold; that proportion is denoted by Ψ, $(1 - \Psi)$ denoting the proportion kept in dollars. Of the latter a proportion γ is kept in the Euro-dollar market. The remaining $(1 - \gamma)$ of dollar reserves is split between a proportion λ held with *U.S.* commercial banks and a proportion $(1 - \lambda)$ held with the Fed.

European commercial banks receive European currency deposits from European residents, as before. These are denoted by M'^*. In addition, they receive dollar deposits (Euro-dollar deposits) from three sources: ED from *U.S.* residents; ED^* from the European public; and CED from Europe's central bank. European commercial banks keep European currency reserves of R^* with their central bank, as a given proportion r^* of their European currency deposits M'^*. They keep dollar reserves of DEB with *U.S.* commercial banks, as a given proportion r_d of their total (Euro-) dollar liabilities. (For a discussion of r_d, see Section IV, below.)

American commercial banks, in addition to dollar deposits by residents M', receive dollar deposits from Europe's central bank and from its commercial banks. They keep reserves in a proportion r to their total deposits. The Fed incurs liabilities to both *U.S.* commercial banks and to the European central bank.

Finally, the public desires to keep a given proportionate relationship between its Euro-dollar and other deposits. American residents keep the ratio of their Euro-dollars (ED) to their total deposits ($M' + ED$) equal to ρ. Similarly, European residents keep a proportion Ω of their total deposits ($M'^* + ED^*$) in the form of Euro-dollar deposits (ED^*). The Euro-dollar deposits held by the nonbank residents of each country are counted as being part of the world money stock in the hands of the residents of that country. Thus, the money stock in the hands of the *U.S.* public is redefined as $M = M' + ED$, that in the hands of Europe's public as $M^* = M'^* + ED^*$, and the world money stock as $M_w = M + M^*$. It is assumed that the demand for M is proportionate to the level of economic activity in the United States, the demand for M^* to that in Europe, and hence, that M and M^* must stand in a relationship β to ensure payments equilibrium, as indicated below:[16]

$$(9) \quad M = \beta M_w = \beta(M + M^*) = \beta(M' + ED + M'^* + ED^*)$$

Keeping these assumptions in mind, it is possible to derive a reduced-form expression for the world money stock. The result for the general case where money multipliers are allowed to differ as between banking systems is quite complicated and is relegated to the Appendix. The already rather formidable result for the special case that abstracts from asymmetries due to differences in reserve ratios kept by commercial banks (i.e., set $r = r^* = r_d = 1/m$) is given below:

$$(10) \quad M_w = m[A + G + A^*\{1 - (1 - \Psi) \cdot [(1 - r)(1 - \gamma)\lambda + (1 - r^2)\gamma]\}] \div [\beta[1 - \rho(1 - r)] + (1 - \beta)\{r\Omega + (1 - \Omega)\{1 - (1 - \Psi)[(1 - r)(1 - \gamma)\lambda + (1 - r^2)\gamma]\}\}]$$

An examination of this expression suggests a number of conclusions that will be illustrated with the help of special cases in subsequent sections of this essay.

Most striking is the fact that the basic symmetry of the gold standard is lost in the more general case. This is evident from the fact that in general the effect of a change in A is different from that of a change in A^*, since the latter is postmultiplied by a constant in expression (10). The symmetry can be regained if either $\Psi = 1$, the gold standard case, or both λ and γ are zero, that is,

[16] Counting all Euro-dollar deposits held by the public as part of the world money supply is clearly inappropriate for some purposes. Moreover, the determinants of the "transactions" demand for Euro-dollars and for other deposits may, in fact, be quite different, in contradiction to what is assumed in our model. For instance, the demand for Euro-dollar deposits by European residents may be a function of the volume of trade or of *U.S.* economic activity and similar considerations may apply to the demand for Euro-dollar deposits by the *U.S.* nonbank public.

if Europe's central bank holds reserves neither with the *U.S.* nor with the European commercial banking system. This, as will become clear subsequently, gives us a clue to the basic reason for asymmetries in the system, namely, that what is low-powered money in one part of the system (for instance, deposits with commercial banks) serves as high-powered money in another part of the system (as part of the sources of the foreign-exchange component of the European monetary base).

Second, whatever their differential impact on the effects of open-market operations according to national origin, various patterns of asset preferences will impinge on the size of the multiplier effects of an open-market operation of given national origin. Differentiating, for instance, dM_w/dA with respect to various behavior parameters, one can trace out some of the effects of a change in the structure of asset preferences. An increase in Euro-dollar deposits reinforces the effect of *U.S* open-market policy whether it originates in a switch by *U S* residents from dollars to Euro-dollars, in a switch by the European public, or in an increase in the European central bank's deposits. A switch by Europe's central bank from deposits with the Fed to deposits with *U S* commercial banks has similar effects as has a decrease in its gold holdings.

III. The Dollar Standard

To understand the origin of the asymmetries noted above, consider a pure dollar standard. That is, assume that the European central bank holds no gold and that there is no Euro-dollar market. This implies that $\Psi = r_d = \rho = \Omega = \gamma = 0$. Also assume for simplicity, as will be done in the remainder of the text, that all reserve ratios of commercial banks are equal. Under these simplifying assumptions, equation (10) becomes:

$$(11) \quad M_w = m \cdot \frac{A + A^*[1 - \lambda(1 - r)]}{\beta + (1 - \beta)[1 - \lambda(1 - r)]}$$

It is immediately apparent that λ, the proportion of dollar reserves held with the *U.S.* commercial banking system, has an important role to play. The higher λ the larger the world money supply multiplier applicable to an increase in A. With $\lambda = 0$, formula (11) becomes:

$$(12) \quad M_w = m \cdot \frac{A + A^*}{\beta + (1 - \beta)} = m(A + A^*)$$

This is exactly the same formula as the gold standard one when r is set equal to r^* and the gold stock is neglected. In other words, a dollar standard where the European central bank holds its reserves with the Fed operates, at least in some respects, exactly like the gold standard. The reason is simply that an open-market operation that leads to inflows or outflows of foreign-exchange reserves has the same impact on the reserves available to commercial banks in the two systems: an outflow of gold from the United States lowers the reserves of American commercial banks by lowering the sources of the money base; an outflow of dollars reduces *U S.* commercial bank reserves by the increase in the European central bank's reserves (dD^*), given A.

With $\lambda = 1$, that is, when Europe's central bank holds all its reserves with *U.S.* commercial banks, formula (11) becomes:

$$(13) \quad M_w = \frac{mA + A^*}{\beta + r(1 - \beta)}$$

An important asymmetry is introduced since an open-market operation in the United States changes the world money stock by m times more than an equal-size open-market operation in Europe. Moreover, the absolute size of dM_w/dA is increased and that of dM_w/dA^* is decreased in comparison with the case where $\lambda = 0$. These asymmetries are explained by the fact that reserve holdings by the European central bank, D_2^*, are a source of the high-powered base of the European money supply while they compete with lower-powered money in the liabilities of the *U.S.* commercial banking system. The reserve outflow created by expansionary monetary policy in the United States diminishes the *U.S.* money stock by less than it increases

TABLE 2

	$\lambda = 0$	$\lambda = 1$
$\frac{dM_w}{dA}$	m	$\frac{m^2}{m\beta + (1-\beta)}$
$\frac{dM_w}{dA^*}$	m	$\frac{m}{m\beta + (1-\beta)}$
$\frac{dM}{dA}$	$m\beta$	$\frac{m^2\beta}{m\beta + (1-\beta)}$
$\frac{dM}{dA^*}$	$m\beta$	$\frac{m\beta}{m\beta + (1-\beta)}$
$\frac{dM^*}{dA}$	$m(1-\beta)$	$\frac{m^2(1-\beta)}{m\beta + (1-\beta)}$
$\frac{dM^*}{dA^*}$	$m(1-\beta)$	$\frac{m(1-\beta)}{m\beta + (1-\beta)}$
$\frac{dIR^*}{dA}$	$(1-\beta)$	$\frac{m(1-\beta)}{m\beta + (1-\beta)}$
$\frac{dIR^*}{dA^*}$	$-\beta$	$\frac{-m\beta}{m\beta + (1-\beta)}$

the European one, the redistribution of foreign-exchange reserves thus increasing the world money stock by $(m - 1)$ times the reserve flow.

The importance of the pattern of reserve holdings by Europe's central bank is illustrated in Table 2, which gives the effect of changes in A and A^* on various variables under the two extreme cases where $\lambda = 0$ and $\lambda = 1$. The $\lambda = 0$ column of this table contains the same elements as the second column of Table 1. This confirms the identity of this version of the dollar standard with the gold standard. The second column of Table 2 is equal to the first column multiplied by $m/(m\beta + (1 - \beta)) > 1$ for the derivatives with respect to A, and by $1/(m\beta + (1 - \beta)) < 1$ for the derivatives with respect to A^*, with the exception of dIR^*/dA^*, this last derivative being multiplied by $m/(m\beta + (1 - \beta))$. This confirms the conclusions reached above that the holding of Europe's foreign-exchange reserves with U.S. commercial banks makes American monetary policy more effective (in the somewhat limited terms of "bang per buck" to be discussed further at the end of this section) and European monetary policy less effective. The counterpart to the loss of effectiveness of European open-market operations in terms of money supply changes is their increased impact on Europe's international reserves. This is a factually relevant conclusion as in practice European central banks have tended to keep few reserves with the Fed and a large proportion of their dollar reserves in U.S. government securities, a custom that has similar effects analytically to keeping them with U.S commercial banks.

As a matter of fact, the practice of keeping European foreign-exchange reserves with the U.S commercial banking system is equivalent to sterilization of $(1 - r)$ of any reserve flow by the Fed. A European reserve gain deposited with U.S. commercial banks diminishes the reserves available to those banks for backing of U.S. resident held dollar balances by rD_2^* given A; had the reserves been deposited with the Fed, the fall in U.S. commercial bank reserves would have been equal to the European gain of foreign-exchange reserves.[17] In other words, the effect on the world money supply of a European reserve gain of one dollar deposited with U.S. commercial banks is the same as that of a one dollar reserve gain deposited with the Fed, $(1 - r)$ of which is neutralized. When Europe's central bank holds its reserves in U.S. Treasury Bills, a European payments surplus exerts no contractionary effect on the U.S. money supply. The money initially lost by the United States is put back into circulation when Europe buys U.S. Treasury Bills. Europe, in effect, performs open-market operations in the United States and neutralizes, as it were, on the Fed's behalf.[18]

[17] It is easily shown that when the Fed neutralizes the D^* but not the D_2^* component of reserve flows, the world money supply formula becomes

$$M_w = m \cdot \frac{\bar{B} + \lambda r A^*}{\beta + (1 - \beta)\lambda r}$$

[18] This result holds, strictly, only in the case of perfect capital mobility (perfect substitutability of U.S and European bonds). Consider an open-market purchase of European bonds by the European central

One may question the relevance of a discussion of the effectiveness of monetary policy in terms of the partial derivative of the world money stock with respect to open-market operations on the grounds (a) that monetary authorities are interested in the national and not the world money stock, and (b) that reduced effectiveness can be compensated by a higher dose of open-market operations. In the model, however, any change in effectiveness with respect to M_w immediately translates into a similar change with respect to M and M^* given β. Second, though bang per buck of open-market operation may not be a matter of great concern in the closed economy, it is a relevant concern for the policymaker in a fixed exchange rate open economy with a finite stock of international reserves (or of domestic assets). For, other things equal, the less "effective" an open-market operation in terms of domestic and world money stocks, the greater will be the loss of international reserves associated with expansionary purchases of bonds.

The final and related point to be made in this section concerns one additional difference between the gold and dollar standards: whereas the given world stock of gold puts a limit on the extent of U.S. monetary expansion compatible with maintenance of the system, no such limit exists, in theory, in the case of the dollar standard.

IV. The Euro-Dollar Market

In the preceding section, no allowance was made for the existence of the Euro-dollar market. The latter impinges on the general model's results in four ways: European residents want to hold a proportion Ω of their money balances as dollar deposits with European commercial banks; *U.S.* residents, likewise, want to keep a proportion ρ of their money holdings with European commercial banks; the European central bank keeps γ of its dollar reserves in the Euro-dollar market; and European commercial banks keep dollar reserves with *U.S.* commercial banks as a fraction r_d of their total dollar deposits.

This last assumption is admittedly *ad hoc*, but so is any assumption of fixed reserve ratios even when minimum ratios are set by law. The assumption can be justified in broad terms on both theoretical and factual grounds. In the first place, European banks do keep demand and time deposits in New York and their volume has been increasing together with the volume of Euro-dollar business though it is impossible to apportion the increase in deposits into reserves against Euro-dollar deposits, regular working balances, etc. Second, some observers and participants in the Euro-dollar market have argued that maturities of assets and liabilities were closely matched, currency by currency, in the Euro-currency market and that little, if any, of the assets would be held in lower-yielding instruments in New York. Though this may be true of interbank deposits within Europe it will not hold for liabilities to and assets on nonbank institutions or the United States, the only net positions appearing in the consolidated balance sheets of this paper, intra-European interbank deposits having been netted out in the balance sheet consolidation process. Holding of reserves in New York against liabilities to nonbanks seems a sensible and prudent practice. Finally, and this is a related point, traditional banking analysis suggests that an individual bank will choose to hold a positive level of reserves even in an unregulated system. Voluntary reserve holdings are a simple consequence of maximization of expected returns in the face of rising costs of illiquidity and a stochastic supply of deposits to the bank; the holding of excess reserves is

---

bank. That bank will have to sell dollar bonds to acquire back its own currency now in excess supply on the foreign-exchange market. In the end, the total amount of bonds outstanding will be the same, the authorities having swapped *U.S.* bonds against European bonds with the public. If the two types of bonds are not perfect substitutes, however, portfolio equilibrium in terms of stocks would imply a fall in the interest rate on European securities and a rise in the rate on dollar assets. As a result, β would tend to fall and the world and European money supply to rise.

indirect evidence for this thesis. As Euro-dollar deposits and withdrawals are typically settled by the transfer of claims on New York banks, the latter constitute the natural reserve instrument for Euro-banks. Though in fact r_d may be less stable than other reserve ratios, it will be assumed to be given here; consequences of its variation can, however, be easily traced in the model.

To analyze the impact of the Euro-dollar market on the world money supply process, it will be convenient to examine three special cases. The Euro-dollar market is grafted in turn on a gold standard, a dollar standard with $\lambda = 0$, and a dollar standard with $\lambda = 1$. We assume for simplicity that $r_d = r = r^*$.[19]

Grafting the Euro-dollar market onto the gold standard implies setting $\Psi = 1$ (the European central bank keeps only gold reserves) in formula (10) above to yield:

$$(14)\quad M_w = m[A + G + A^*] \div [\beta[1 - \rho \cdot (1 - r)] + (1 - \beta)[1 - \Omega(1 - r)]]$$

Expression (14) indicates that the system is still symmetrical with respect to equal changes in the various components of the base (A, A^*, and G). The reason is that reserve flows set in motion increases in the supply of money in the hands of residents of the surplus country that are exactly offset by decreases in the deficit country. The world money supply multiplier, however, is larger than what it would be without a Euro-dollar market. The reason is simply that European commercial banks keep reserves against their Euro-dollar deposits with *U.S. commercial* banks and not with a central bank. Again, what is high-powered money in Europe (DEB) is low-powered money from the point of view of the *U.S.* banking system.

Consider now the second special case mentioned above by setting $\Psi = 0$ and $\lambda = 0$. The resulting reduced-form equation for M_w is

$$(15)\quad M_w = m \cdot [A + A^*\{1 - (1 - r^2)\gamma\}] \div [\beta[1 - \rho(1 - r)] + (1 - \beta) \cdot \{r\Omega + (1 - \Omega)[1 - (1 - r^2)\gamma]\}]$$

Central bank holdings of dollars in the Euro-dollar market, in a proportion γ to their dollar holdings with the Fed, introduce a new dimension into the system. Asymmetries arise and dM_w/dA increases while the derivative of M_w with respect to A^* decreases. The explanation again arises from the fact that a loss of foreign-exchange reserves by the United States creates a smaller decrease in commercial bank reserves there than an equal gain of foreign-exchange reserves expands commercial bank reserves in Europe. As the European central bank's foreign-exchange reserves expand, part of the gain is deposited in European commercial banks, which in turn redeposit a fraction r_d of this increase in their dollar liabilities (CED) with *U.S.* commercial banks.

A further multiplicative element is added when the European *central* bank keeps a fraction λ of its foreign exchange reserves with *U.S.* commercial banks. For the case where $\lambda = 1$, the reduced-form equation for M_w becomes:

$$(16)\quad M_w = m \cdot [A + A^*\{1 - [(1 - r) \cdot (1 - \gamma) + (1 - r^2)\gamma]\}] \div [\beta[1 - \rho \cdot (1 - r)] + (1 - \beta)\{r\Omega + (1 - \Omega) \cdot \{1 - [(1 - r)(1 - \gamma) + (1 - r^2)\gamma]\}\}]$$

Positive official dollar holdings with *U.S* commercial banks D_2^* adds a multiplier effect that is similar to that described for the special case $\lambda = 1$ in the section discussing the dollar standard.[20]

In conclusion, a switch from traditional national currency holdings to Euro-dollar deposits, be it by the European public, the *U.S.* public, or the European central bank, tends to expand the world money supply,

[19] The assumption that r_d is equal to the other two reserve ratios is purely a matter of convenience If it is, as is perhaps likely, lower than either r or r^*, the analysis can be carried out without important qualitative changes with the help of the general reduced form for the world money stock given in the Appendix

[20] When the Fed neutralizes all reserve flows, the world money supply becomes $M_w = m\bar{B}/\beta(1 - \rho)$

TABLE 3

	$\frac{dM_w}{dA}$	$\frac{dM_w}{dA^*}$	$\frac{dIR^*}{dA}$	$\frac{dIR^*}{dA^*}$
Pure gold or dollar standard				
1. $\lambda = \rho = \Omega = \gamma = 0$	4(8)	4(0)	.5(1)	–.5(–1)
Pure dollar standard				
2. $\lambda = .5, \Psi = \Omega = \rho = \gamma = 0$	4.9(7.1)	3.1(.89)	.62(.89)	–.62(–.89)
3. except $\lambda = .8$	5.7(6.7)	2.3(1.3)	.71(.83)	–.71(–.83)
4. except $\lambda = 1$	6.4(6.4)	1.6(1.6)	.8(.8)	–.8(–.8)
Euro-dollars and gold standard				
5. $\Psi = 1, \gamma = 0$	4.8	4.8	45	–.55
Euro-dollars and dollar standard				
6. $\lambda = 0, \Psi = 0$	5.3	4.3	.49	–.60
7. except $\lambda = 1$	7.5	1.6	.70	–.85
General case				
8. See note (a)	5.9(10)	3.5(0)	.55(.94)	–.67(–1)

Notes: (a) Unless otherwise indicated assumed values of behavior parameters are as follows: $\beta = .5, r = .25, \Psi = .4, \lambda = .8, \Omega = .25, \gamma = .2, \rho = .2$.

(b) Numbers in parentheses indicate results for the case of neutralization. In rows 1 and 8 all reserve flows are neutralized; in rows 2, 3, and 4, only European dollar reserves held at the Fed are neutralized. The derivatives in the neutralization case are to be interpreted as being taken with respect to $\bar{B}$.

(c) $\frac{dIR^*}{dA} = r(1-\Omega)(1-\beta)\frac{dM_w}{dA}, \quad \frac{dIR^*}{dA^*} = r(1-\Omega)(1-\beta)\frac{dM_w}{dA^*} - 1$

other things equal, and creates or reinforces asymmetries that increase the effectiveness of *U.S.* monetary policy and decrease that of European monetary policy.[21] Note, finally, that to derive a Euro-dollar multiplier formula would violate the spirit of the general method of analysis used in this paper, for such a formula must assume exogenous flows of dollar reserves to the Euro-dollar market. In the present analysis, these flows are endogenous, and the effect of autonomous changes in monetary policy or asset preferences on the Euro-dollar market can be studied instead.

V. A Numerical Example

The relevance of the institutional arrangements outlined above to the long-run impact of national monetary policy and to the functioning of the international monetary system can be highlighted with the help of the numerical examples shown in Table 3.

It is assumed throughout the table that the United States and the rest of the fixed exchange rate world are of roughly equal economic size (i.e., that $\beta = .5$) and that reserve ratios are the same in the two countries and in the Euro-dollar market and equal to .25 (i.e., $m = 4$). Row 1 in Table 3

[21] The hedging statement "tends" in the text has been inserted to take into account an ambiguity in some of the partial derivatives of expression (10) with respect to a number of parameters. However, imposing the restriction that European reserves be positive (i.e., that $A > \beta/(1-\beta)A^*$), enables one to establish that $\partial M_w/\partial\Psi$, $\partial M_w/\partial r$, $\partial M_w/\partial\gamma < 0$, and that $\partial M_w/\partial\lambda > 0$. In addition, $\partial M_w/\partial\rho > 0$. Finally, $\partial M_w/\partial\Omega$ will tend to be positive, the smaller γ and λ. More precisely, positivity of $\partial M_w/\partial\Omega$ requires that $1 - [(1-\lambda)\cdot\gamma + \gamma] > r\gamma$, that is, that the ratio of European central-bank reserves held with the Fed (or in gold) must be greater than that held in Euro-banks times r. The intuitive explanation is as follows. Suppose Europe only holds Euro-dollars (CED) as international reserves. A switch by European residents towards Euro-dollar deposits increases European commercial banks' demand for dollar deposits in New York, but these can only be made available through a decrease in the deposits available to private *U.S.* holders, that is, through a decrease in the *U.S.* money stock M, if the *U.S.* monetary base is given as it will be if the switch does not result in a reduction of the Fed's liabilities to foreign official holders or in an increase in its gold stock.

illustrates the simplest gold or pure dollar standard where Europe's central bank holds all its reserves with the Fed. A $1 open-market operation by either Europe or the United States increases the world money stock by four times as much and each national money stock by half that amount (since $\beta = .5$); a gain in reserves of .5 brings about the required increase of 2 in the surplus country's money stock, given the national money multiplier of 4. The results for the case where the United States sterilizes all reserve flows under the gold or simplest dollar standard case are given in parenthesis in row 1. As stated in previous sections, neutralization by the United States reduces the effectiveness of Europe's monetary policy with respect to its effect on money supplies to zero, but makes European open-market operations an extremely efficient means of controlling its stock of reserves. Since $\beta = .5$, the effectiveness of *U.S.* monetary policy is doubled (remembering that $dM_w/d\bar{B} = m/\beta$), an increase of $1 in the *U.S.* monetary base means that the *U.S.* money stock has to increase by m dollars in equilibrium. This is only possible if Europe's money stock increases by $(1 - \beta)/\beta$ times this amount, or, in our example by $4. This will have occurred when Europe has gained $1 of reserves and total domestic assets held by the Fed have increased by $2.

Consider now row 8 of Table 3. The results there are based on the assumptions that 40 percent of Europe's reserves are held in the form of gold ($\Psi = .4$); that 20 percent of the remainder is held in the Euro-dollar market ($\gamma = .2$); that, of European reserves held in the United States, 80 percent are deposited with *U.S.* commercial banks and 20 percent with the Fed ($\lambda = .8$); and that the European and American public keep 25 and 20 percent, respectively, of their total money holdings in the Euro-dollar market ($\Omega = .25$ and $\rho = .2$). These are not entirely unrealistic assumptions as to the magnitudes involved in the mid-1960's, though they perhaps overestimate the role of the Euro-dollar market (at least in terms of average if not in terms of marginal ratios) and underestimate that of the dollar standard and neutralization.[22] The result, as compared with row 1, is an increase in the "effectiveness" of *U.S.* monetary policy from 4 to 5.9 and a decrease in that of European monetary policy from 4 to 3.5.

To gain some feeling of the respective importance of the dollar standard, neutralization, and the Euro-dollar market in influencing various measures of the impact of monetary policy, consider briefly rows 2 through 7 of Table 3. Comparing rows 1 and 5, it immediately appears that, even on our perhaps exaggerated assumptions about the actual importance of the Euro-dollar market, the increase in the world money supply multiplier entailed by the market, though significant, is not huge. Furthermore, as long as the European central bank does not hold reserves in the Euro-dollar market, symmetry with respect to the effect of open-market operations on the world money stock, though not on foreign-exchange reserves, prevails. The asymmetry with respect to reserve changes arises from the fact that an increase in the world money stock brought about by an increase in A or A^* increases the demand for dollars by European residents, and hence absorbs part of the initial excess supply of *dollars* dA, but not of European currency dA^*. In contrast to the case where $\gamma = 0$, holdings of Euro-dollar deposits by the European central bank ($\gamma = .2$) introduce asymmetries in the impact of monetary policy, raising the effectiveness of *U.S.* policy and lowering that of European monetary policy, as indicated by a comparison of rows 5 and 6.

Clearly, however, the strongest asymmetries arise from the practice of holding foreign-exchange reserves with *U.S.* commercial banks (compare cases where λ takes on increasingly higher values) and from

[22]The results in Genberg and the author suggest that the behavior of the world money stock in the period 1957–71 is not inconsistent with the hypothesis of complete neutralization by the United States (and a zero long-run multiplier for European open-market operations). The world money stock in that paper does not include Euro-dollar deposits.

neutralization by the United States, as perusal of Table 3 immediately indicates.[23]

VI. Conclusions

The method of analysis outlined in this paper seems appropriate to the investigation of a number of issues. It does emphasize that, under strictly fixed exchange rates, international monetary theory can make use of the concepts developed for closed-economy monetary theory, adding to it distributional considerations effected through the payments adjustment mechanism. It also shows that the specific institutional pattern ruling interbank and intercountry relations has an important impact on the effectiveness of monetary policy at both a global and national level.

The method of analysis, however, suffers from the same shortcomings as does closed-economy money multiplier analysis when applied to problems with which it is not equipped to deal. This suggests a number of extensions. First, models of the world money supply could be integrated with an explicit general equilibrium analysis of the determinants of output and interest rate fluctuations. This would be particularly appropriate for an analysis of the short run before all variables have fully adjusted. Second, and again in a short-run context, the dynamics of the money supply process could be analyzed by formulating stock adjustment functions.[24]

Finally, empirical investigation of the world money supply process is needed to assess the importance of the asymmetries and multiplier effects emphasized above as well as to gain a better understanding of the importance of the institutional changes that have occurred in the world's fixed exchange rate system from its gold standard heyday to its recent and partial demise. The method of analysis proposed in this paper is designed to provide a tentative framework for such an investigation.[25]

APPENDIX

1) Recall the *balance sheet structure* of the general model of Section II:

United States

Fed Assets	Fed Liabilities	Commercial Banks Assets	Commercial Banks Liabilities
A	R	R	M'
αG	D^*	L_1	D_2^*
			DEB

Europe

Central Bank Assets	Central Bank Liabilities	Commercial Banks Assets	Commercial Banks Liabilities
A^*	R^*	R^*	M'^*
$(1-\alpha)G$		L_1^*	ED^*
D^*		DEB	ED
D_2^*			CED
CED			

2) *Definition of behavior parameters:*

U.S. reserve ratio against commercial banks domestic currency liabilities:

$$r = \frac{1}{m} = \frac{R}{M' + D_2^* + DEB}$$

European reserve ratio against commercial banks domestic currency liabilities:

$$r^* = \frac{1}{m^*} = \frac{R^*}{M'^*}$$

European commercial banks dollar reserve ratio against Euro-dollar liabilities:

[23] Comparing rows 1, 2, 3, and 4, note that asymmetries introduced by neutralization diminish as λ rises. This is due to the fact that holding dollar reserves with U.S. commercial banks is equivalent to neutralization on behalf of the United States and, hence, the less the total increase in A needed to sustain a given increase in $\bar{B}$. When $\lambda = 1$, the required neutralization by the United States is zero as indicated by the equality of the original results with those in parentheses in row 4.

[24] It may be worthwhile to note that the money supply formulae developed throughout the text incorporate elements of money demand functions since β is in equilibrium equal to the ratio of the first country's demand for money to the sum of the demands for money in the world.

[25] For a first attempt at econometric modelling along the lines suggested in the text, see Genberg and the author.

$$r_d = \frac{1}{m_d} = \frac{DEB}{ED^* + ED + CED}$$

Proportion of European residents' money holdings held as Euro-dollars:

$$\Omega = \frac{ED^*}{M^*} = \frac{ED^*}{M'^* + ED^*}$$

Proportion of *U.S.* residents' money holdings held as Euro-dollars:

$$\rho = \frac{ED}{M} = \frac{ED}{M' + ED}$$

Equilibrium proportion of world money stock held by residents of country 1:

$$\beta = \frac{M}{M_w} = \frac{M' + ED}{M_w}$$

Proportion of Europe's foreign-exchange reserves held in gold:

$$\Psi = \frac{(1-\alpha)G}{R^* - A^*}$$

Proportion of Europe's dollar reserves held in Euro-dollar market:

$$\gamma = \frac{CED}{D^* + D_2^* + CED}$$

Proportion of Europe's reserves in the United States held with *U.S.* commercial banks:

$$\lambda = \frac{D_2^*}{D^* + D_2^*}$$

Proportion of *U.S.* gold holdings in the (given) world gold stock (G), an *endogenous* variable:

$$\alpha G$$

3) *Definition of money stocks:*

1. $M = M' + ED$
2. $M^* = M'^* + ED^*$
3. $M_w = M + M^* = M' + M'^* + ED + ED^*$

4) *Derivation of general reduced form for the world money stock:*

The behavior parameters and definitions of the money stock are used to derive a world money supply formula. The most important relations used in the process are given below:

1. $G = \alpha G + (1 - \alpha)G$
2. $R = r\{M' + D_2^* + DEB\} = \frac{1}{m} \cdot \{M' + D_2^* + DEB\}$
3. $R^* - A^* = \frac{1}{\Psi}(1-\alpha)G = \frac{1}{1-\psi} \cdot \{D^* + D_2^* + CED\}$
4. $R^* = r^* M'^* = \frac{1}{m^*} M'^*$
5. $DEB = r_d[ED^* + ED + CED] = \frac{1}{m_d} \cdot [ED^* + ED + CED]$
6. $ED^* = \Omega M^* = \Omega(M'^* + ED^*)$
7. $ED = \rho M = \rho(M' + ED)$
8. $M = \beta M_w$
9. $M^* = (1 - \beta)M_w$

Through a tedious process of substitution the following general reduced term for M_w is obtained:[26]

$$M_w = [A + G + A^*\{1 - (1-\Psi)[(1-r) \cdot (1-\gamma)\lambda + \gamma(1 - rr_d)]\}] \div [\beta\{r[1 - \rho(1 - r_d)]\} + (1-\beta) \cdot \{rr_d\Omega + r^*(1-\Omega)\{1 - (1-\Psi) \cdot [(1-r)(1-\gamma)\lambda + (1 - rr_d)\gamma]\}\}]$$

[26] To illustrate the process of substitution, take the derivation of equation (8) in the text, the reduced-form expression for M_w relevant to the simple gold standard case discussed in Section I. From equation (7) recall that

$$M_w = M + M^* = mA + m^*A^* + m^*G + (m - m^*)\alpha G$$

Note also that the gold reserves of the first country can be expressed as

$$\alpha G = rM - A = r\beta M_w - A$$

Substitute the second expression into the first and simplify to obtain

$$M_w = m^*(A + A^* + G) + (m - m^*)r\beta M_w$$

Solve for M_w and divide numerator and denominator by m^* to obtain expression (8) in the text.

REFERENCES

K. Brunner and A. Meltzer, "Some Further Investigations of the Demand and Supply Functions of Money," *J. Finance*, May 1964, *19*, 240 83.

Philip Cagan, *Determinants and Effects of Changes in the Stock of Money, 1875 1960*, New York 1965

Milton Friedman and Anna Schwartz, *A Monetary History of the United States, 1867 1960*, Princeton 1963.

H. Genberg and A. K. Swoboda, "Worldwide Inflation under the Dollar Standard," GIIS-Ford disc. paper no. 12, Grad. Instit. Int. Stud., Geneva, Jan. 1977.

P. J. K. Kouri and M. G. Porter, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May/June 1974, *82*, 443 67.

James E. Meade, "The Amount of Money and the Banking System," *Econ. J.*, 1934, *44*, 77 83; reprinted in Friedrich A. Lutz and Lloyd W. Mints, eds., *Readings in Monetary Theory*, London 1951.

———, *The Balance of Payments*, London 1951.

R. A. Mundell, "The International Disequilibrium System," *Kyklos*, 1961, No. 2, *14*, 153-70.

A. K. Swoboda, "Eurodollars and the World Money Supply: Implications and Control," in his *Europe and the Evolution of the International Monetary System*, Leiden; Geneva 1973, ch. 10.

——— **and R. Dornbusch,** "Adjustment, Policy, and Monetary Equilibrium in a Two-Country Model," in Michael B. Connolly and Alexander K. Swoboda, eds., *International Trade and Money*, London 1973, 225-61

Equilibrium in an Imperfect Market: A Constraint on the Number of Securities in the Portfolio

By HAIM LEVY*

The pioneering work of Harry Markowitz (1952, 1959) and James Tobin in portfolio theory has led to the development of a theory of the pricing of capital assets under uncertainty. This theory, well-known in the literature as the capital asset pricing model (*CAPM*), was developed independently by William Sharpe, John Lintner (1965a), and Jack Treynor. Two basic related properties implied by the *CAPM* are: (a) that all investors hold in their portfolio all the risky securities available in the market, and (b) that investors hold the risky assets in the same proportions, as these assets are available in the market, independent of the investors' preference.[1] This latter property of the *CAPM* makes it possible to draw many conclusions regarding the equilibrium risk-return relationship of risky assets.

Properties (a) and (b) contradict the market experience as established in all empirical research. First, investors differ in their investment strategy and do not necessarily adhere to the same risky portfolio. Second, the typical investor usually does not hold many risky assets in his portfolio. Indeed, in a recent study, Marshall Blume, Jean Crockett, and Irwin Friend found that, in the tax year 1971, individuals held highly undiversified portfolios. The sample, which included 17,056 individual income tax forms, revealed that 34.1 percent held only one stock paying dividends, 50 percent listed no more than two, and only 10.7 percent listed more than ten. Though only firms paying cash dividends were included in this statistic, it is obvious that most individuals held a relatively small number of stocks in their portfolio. Another source of data which confirms these findings is the Federal Reserve Board's 1967 survey of the Financial Characteristics of Consumers. This survey covered all households whether or not they filed income tax forms. According to this survey, the average number of securities in the portfolio was 3.41.[2]

The fact that properties (a) and (b) do not conform to reality is not a sufficient cause for rejecting the theoretical results of the *CAPM* One could also accept the *CAPM* results on positive grounds. If the theoretical model does indeed explain the price behavior of risky assets, one could argue that investors behave *as if* properties (a) and (b) were true, in spite of the fact that these properties obviously do not prevail in the market. Unfortunately, we can *not* justify the theoretical results of the *CAPM* on positive grounds.

To illustrate the latter difficulty, let us return in greater detail to the *CAPM*. According to the *CAPM*, the expected return on asset i, $E(R_i)$ is related to the expected return on the market portfolio $E(R_m)$ as follows:

$$E(R_i) - r = [E(R_m) - r]\beta_i \tag{1}$$

where r is the risk-free interest rate, β_i is the risk index of the ith security (the "systematic risk") and is defined as $Cov(R_i, R_m)/Var(R_m)$, and R_m is the rate of return on a portfolio which consists of all available risky assets and is called the "market portfolio."

Although the *CAPM* is formulated in terms of *ex ante* parameters, it is common to employ *ex post* data rather than *ex ante* values in empirical studies. Thus, we first

*Hebrew University of Jerusalem. I acknowledge the helpful comments of Yoram Landskroner, Yoram Kroll and an anonymous referee of this *Review*.

[1]Lintner (1969) extends the *CAPM* to the case of disagreement of investors with regards to expected parameters. I assume in this model that investors agree with regard to future parameters but the model presented in this paper can be easily extended to the case of disagreement.

[2]For more details of these findings and their analysis, see Blume and Friend (1975)

run a time-series regression,

$$(2) \qquad R_{it} = \alpha_i + \beta_i R_{mt} + e_{it}$$

and estimate the systematic risk $\hat{\beta}_i$ of each asset i (where R_{it} and R_{mt} are the rates of return of the ith asset and the market portfolio, respectively, in year t). In the second step, in order to examine the validity of the *CAPM*, we run a cross-section regression,

$$(1') \qquad \bar{R}_i - r = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\beta}_i + u_i$$

where $\bar{R}_i$ is the average return on the ith risky asset, $\hat{\beta}_i$ is the estimate of the ith asset systematic risk, taken from the time-series regression, and u_i is a residual term. If the *CAPM* is valid one should obtain (see equation (1)) in equilibrium, $\hat{\gamma}_0 = 0$ and $\hat{\gamma}_1 = \bar{R}_m - r$, where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are the regression coefficients estimated by (1'), and $\bar{R}_m$ is the average observed rate of return on the market portfolio (for example, average rate of return on Standard and Poor's index).

Unfortunately, in virtually all empirical research,[3] it emerges that $\hat{\gamma}_0$ is significantly positive and $\hat{\gamma}_1$ is much below $\bar{R}_m - r$. For rates of return of *individual* stocks the correlation coefficient of (1') is very low if one employs monthly rates of return, and only 20-25 percent with annual rates of return.[4]

Finally, in virtually all empirical studies, formulation (3) increases the correlation coefficient,

$$(3) \qquad \bar{R}_i - r = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\beta}_i + \hat{\gamma}_2 \hat{S}^2_{e_i}$$

where i stands for the ith security and $\hat{S}^2_{e_i}$ is the residual variance around the time-series regression (2), i.e., the variance of the residuals e_{it}. In this formulation the estimate γ_2 happens to be significantly positive, contrary to the expected results from the *CAPM* since, if the *CAPM* is correct, one should find that $\gamma_2 = 0$. Moreover, in most cases, the contribution of $\hat{S}^2_{e_i}$ to the coefficient of correlation is even more important than the contribution of the systematic risk, $\hat{\beta}_i$.

In this paper I try to narrow the gap between the theoretical model and the empirical findings by deriving a new version of the *CAPM* in which investors are assumed to hold in their portfolios some given number of securities. Obviously, investors' portfolios differ in the proportions of risky assets and even in the types of risky assets that they hold. This, of course, is consistent with investors' behavior as established in previous empirical research. I denote the modified model as *GCAPM* (general capital asset pricing model), since the *CAPM* emerges as a special case.

The derivation of the *GCAPM* under these conditions is given in Section II. In the third section I show that the modified model explains the discrepancy between the theoretical results of the *CAPM* and the empirical findings mentioned above. Some empirical results are presented which confirm that the systematic risk β_i plays no role in explaining price behavior, once the variance is taken into account, (Section IV). Concluding remarks are given in Section V.

I. Equilibrium in an Imperfect Market: The *GCAPM*

William Sharpe and Lintner (1965a) have shown that, if there is no constraint on the number of securities to be included in the investors' portfolio, all investors will hold some combination of m, the market portfolio of risky assets, and the riskless asset bearing interest rate r (see Figure 1).

Now, suppose that, as a result of transaction costs, indivisibility of investment, or even the cost of keeping track of the new financial development of all securities, the kth investor decides to invest only in n_k securities. Under this constraint he will have some interior efficient set (of risky assets), say, $A'B'$, and the investor will divide his portfolio between some risky portfolio k

[3]See Fisher Black, Michael Jensen, and Myron Scholes; George Douglas; Lintner (1965b), Merton Miller and Scholes.

[4]I emphasize that the low correlation is obtained when equation (1') is regressed using individual stock. In order to minimize the measurement errors, it is common to use in (1') portfolios rather than individual stocks. This portfolio technique increases the correlation coefficient dramatically However, in spite of the possible errors, individual stocks should be used since the *CAPM* defines equilibrium prices of individual stocks.

FIGURE 1

FIGURE 2

and the riskless asset. Obviously, the investor's welfare will decrease if no more than n_k securities may be included in the portfolio, since for a given expected return, he will be exposed to higher risk (see Figure 1).

In the specific case in which all investors hold the same number of risky assets n_k in equilibrium, all these interior efficient sets will be tangent to the same straight line. To illustrate, suppose that $n_k = 2$ for all k and that there are $n = 3$ risky assets available in the market. Figure 2 shows this possibility using A, B, and C to indicate the three risky securities.

Without any constraints, all investors hold portfolio m (i.e., the market portfolio), and all efficient portfolios lie on line rmM. Now suppose that all investors decide to include only two risky assets in their portfolio. Investors who hold securities A and B are faced with opportunity line rkK. If all investors decide to include two risky assets in their portfolio, this situation will not represent an equilibrium situation, since no one will purchase security C (see Figure 2). Hence the price of security C will decline, and its expected return will increase, until we get a new efficient curve between B and C (or C and A) which will be tangent to line rkK. In this case, however, the market may be cleared out. Note that not all two securities' efficient sets need to be tangent to the market line rkK. A sufficient condition for the market to be cleared out, in this example, is for two out of three efficient sets given in Figure 2 (i.e., AB, BC, AC) to be tangent to the line rkK. In other words, each of the three assets must be included in some two-asset portfolio which is tangent to the straight line.

In the more realistic case, which will be dealt with below, the kth investor has the constraint of investing in no more than n_k risky assets when n_k varies among investors

FIGURE 3

mainly as a function of the size of their wealth. In this case there are many interior efficient sets (see Figure 3), and the existence of many market lines does not contradict the possibility that the market may be in equilibrium.

In this case, rm is the opportunity line without any constraint on the number of securities in the portfolio, $r2$ is the market line with the constraint that no more than two securities are included in the portfolio; $r3$ is the line with the constraint of no more than three securities in the portfolio, etc. Obviously, the same security may be held in proportion of 20 percent of one portfolio, 5 percent of a second portfolio, etc. We derive below the equilibrium prices of risky assets for the general case in which the constraint on n_k varies from investor to investor. Again, a necessary condition for equilibrium in the stock market is that each security be included in at least one of the chosen unlevered portfolios from the above efficient sets.

Let us turn now to the derivation of the risk-return relationship under the constraint that not all risky assets are held in the investors' portfolio. We assume that there are K investors (or groups of investors), and the kth investor wealth is T_k dollars. Furthermore, assume that the kth investor invests only in n_k risky assets while there are in the market $n > n_k$ risky assets. Thus, the kth investor minimizes the portfolio's variance subject to the constraint that the number of securities in his portfolio cannot exceed n_k. More specifically, one has to differentiate partially with respect to x_{ik} and λ_k the Lagrangian function

$$L = \sum_{i=1}^{n_k} x_{ik}^2\sigma_i^2 + 2\sum_{\substack{i=1\\ j>i}}^{n_k} x_{ik}x_{jk}\sigma_{ij} + 2\lambda_k\left[\mu_k - \sum_{i=1}^{n_k} x_{ik}\mu_i - \left(1 - \sum_{i=1}^{n_k} x_{ik}\right)r\right]$$

subject to the constraint that no more than n_k securities will be included in the optimal portfolio, where

σ_i^2 = the variance of the ith security return (per \$1 of investment)

σ_{ij} = the covariance between returns of securities i and j

μ_k = the portfolio expected return

x_{ik} = the proportion invested in the ith security by the kth investor

r = riskless interest rate

λ_k = Lagrange multiplier appropriate for the kth investor

Suppose that the investor selects n_k assets out of the n available assets to be included in his optimal portfolio. Then by differentiating the Lagrangian function we obtain the following $n_k + 1$ equations, which provide the optimal diversification strategy among the n_k risky assets

$$(4)\quad \begin{aligned} x_{1k}\sigma_1^2 + \sum_{j=2}^{n_k} x_{jk}\sigma_{1j} &= \lambda_k(\mu_1 - r)\\ x_{2k}\sigma_2^2 + \sum_{\substack{j=1\\ j\neq 2}}^{n_k} x_{jk}\sigma_{2j} &= \lambda_k(\mu_2 - r)\\ \vdots\qquad\qquad\vdots\qquad &\qquad\vdots\\ x_{nk}\sigma_{n_k}^2 + \sum_{\substack{j=1\\ j\neq n_k}}^{n_k} x_{jk}\sigma_{n_k j} &= \lambda_k(\mu_{n_k} - r)\\ \mu_k = \sum_{i=1}^{n_k} x_{ik}\mu_i &+ \left(1 - \sum_{i=1}^{n_k} x_{ik}\ r\right) \end{aligned}$$

Thus, the optimal investment strategy of the kth investor is given by the vector x_{1k}, $x_{2k}, \ldots, x_{nk}$ which solves the above equations. We multiply the first equation by x_{1k}, the second equation by x_{2k}, etc., and then sum up the first n_k equations to obtain

$$\sigma_k^2 = \lambda_k\left(\sum_{i=1}^{n_k} x_{ik}\mu_i - \sum_{i=1}^{n_k} x_{ik}r\right) = \lambda_k\left[\sum_{i=1}^{n_k} x_{ik}\mu_i + \left(1 - \sum_{i=1}^{n_k} x_{ik}\right)r - r\right] = \lambda_k(\mu_k - r)$$

Hence,

$$(5)\qquad \frac{1}{\lambda_k} = \frac{\mu_k - r}{\sigma_k^2}$$

where μ_k and σ_k^2 are the expected return and variance of the kth investor's optimal portfolio. Using (4) and (5) the kth investor will be in equilibrium if and only if

$$(6) \qquad \mu_i = r + \frac{\mu_k - r}{\sigma_k^2} cov(R_i R_k)$$

where R_i and R_k are the rates of return on the ith security and on the portfolio chosen by the kth investor. Equation (6) can be rewritten as

$$(6') \qquad \mu_i = r + (\mu_k - r)\beta_{k_i}$$

where β_{k_i} is the systematic risk of the ith asset in the kth investor's optimal portfolio R_k and is defined as $\beta_{k_i} = Cov(R_i, R_k)/\sigma_k^2$.

It is important to note that the equilibrium relationship given in equations (6) and (6') is independent of the borrowing or lending policy of the kth investor.[5] Thus, without loss of generality, we can assume that

$$\sum_{i=1}^{n_k} x_{ik} = 1$$

and this will not affect the solution of the optimal investment. In the rest of the paper we assume that μ_k and σ_k^2 are the parameters of the optimal *unlevered* portfolio chosen by the kth investor. This is tantamount to the assumption that

$$\sum_{i=1}^{n_k} x_{ik} = 1$$

In order to examine the impact on equilibrium price determination, of not holding all assets in the portfolio we need to use some algebra. Since $R_k = \Sigma_{j=1}^{n_k} x_{jk} R_j$, equation (6) can be rewritten as

$$(7) \quad \frac{v_{i1} - v_{i0}}{v_{i0}} = r + \frac{(\mu_k - r)}{\sigma_k^2} \cdot \left[x_{ik}\sigma_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} x_{jk}\sigma_{ij} \right]$$

when v_{i1} and v_{i0} stand for the expected market value of firm i at the end of the period, and for the equilibrium present value, respectively. Hence,

$$(8) \quad v_{i1} - v_{i0}(1 + r) = \frac{(\mu_k - r)}{\sigma_k^2} \cdot \left[v_{i0}x_{ik}\sigma_i^2 + v_{i0} \sum_{\substack{j=1 \\ j \neq i}}^{n_k} x_{jk}\sigma_{ij} \right]$$

Let us denote

σ_i^{*2} = the expected variance of the return on *one share* of the ith firm at the end of the investment period

σ_{ij}^* = the expected covariance of the return of *a share* of firm i and *a share* of firm j

N_i = the number of outstanding shares of firm i

P_{i0} = the equilibrium price of a share of firm i

P_{i1} = the expected price of a share of firm i at the end of the period

[5]To be more specific suppose that an investor who owns T_k dollars decides to borrow or lend $(\Sigma_i^{n_k} x_{ik} - 1)$ per each dollar that he owns. Then, if, R_k is the return (per one dollar) on his optimal portfolio solely from risky assets, the return on his selected portfolio (including the borrowing or lending) denoted by R_k^* will be

$$R_k^* = \left(\sum_{i=1}^{n_k} x_{ik}\right) R_k - \left(\sum_{i=1}^{n_k} x_{ik} - 1\right) r$$

and hence

$$\mu_k^* = \left(\sum_{i=1}^{n_k} x_{ik}\right)\mu_k - \left(\sum_{i=1}^{n_k} x_{ik}\right) r + r,$$

$$\sigma_k^{2*} = \left[\sum_{i=1}^{n_k} x_{ik}\right]^2 \sigma_k^2$$

and, $\quad cov^*(R_i R_k) = \left[\left(\sum_{i=1}^{n_k} x_{ik}\right)\right] cov(R_i R_k)$

Rewriting (6) in terms of R_k^* we obtain

$$\mu_i = r + \frac{\sigma_k^* - r}{\sigma_k^{2*}} cov^*(R_i R_k)$$

or

$$\mu_i = r + \frac{\sum_{i=1}^{n_k} x_{ik}(\mu_k - r) + r - r}{\left(\sum_{i=1}^{n_k} x_{ik}\right)^2 \sigma_k^2} \cdot \left(\sum_{i=1}^{n_k} x_{ik}\right) cov(R_i R_k)$$

and finally

$$\mu_i = r + \frac{\mu_k - r}{\sigma_k^2} cov(R_i R_k)$$

where μ_k and σ_k^2 are the expected return and variance of the optimal portfolio of the kth investor when he neither borrows nor lends money

Thus,

$$\sigma_i^{*2} = \sigma_i^2 P_{i0}^2, \ \sigma_{ij}^* = \sigma_{ij} P_{i0} P_{j0}$$

and equation (8) can be rewritten in terms of market price per share,

$$(9) \quad N_i P_{i1} - N_i P_{i0}(1 + r) = \frac{(\mu_k - r)}{\sigma_k^2} \cdot \left[N_i P_{i0} x_{ik} \sigma_i^2 + N_i P_{i0} \sum_{\substack{j=1 \\ j \neq i}}^{n_k} x_{jk} \sigma_{ij} \right]$$

Dividing by N_i yields

$$(10) \quad P_{i1} - P_{i0}(1 + r) = \frac{(\mu_k - r)}{\sigma_k^2} \cdot \left[P_{i0} x_{ik} \sigma_i^2 + P_{i0} \sum_{\substack{j=1 \\ j \neq i}}^{n_k} x_{jk} \sigma_{ij} \right]$$

Now recall that the proportions invested by the kth investor x_{ik} and x_{jk} in the ith and jth assets, respectively, have been given by $x_{ik} = N_{ik} P_{i0}/T_k$, and $x_{jk} = N_{jk} P_{j0}/T_k$, where N_{ik} and N_{jk} stand for the number of shares of firm i and j in the kth investor's portfolio, and T_k is the total amount of dollars invested by him in risky assets. Thus, the substitution of x_{ik} and x_{jk} in equation (10) yields,

$$(11) \quad P_{i1} - P_{i0}(1 + r) = \frac{(\mu_k - r)}{T_k \sigma_k^2} \cdot \left[P_{i0}^2 N_{ik} \sigma_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} N_{jk} P_{i0} P_{j0} \sigma_{ij} \right]$$

By substituting for σ_i^* and σ_{ij}^* (variance and covariances in terms of one share rather than one dollar), and multiplying and dividing by T_k, we obtain,

$$(12) \quad P_{i1} - P_{i0}(1 + r) = \frac{T_k(\mu_k - r)}{T_k^2 \sigma_k^2} \cdot \left[N_{ik} \sigma_i^{*2} + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} N_{jk} \sigma_{ij}^* \right]$$

Equation (12) should apply to the kth investor, but *only for securities which are included in his portfolio.*

Now, in order to have price equilibrium in terms of the aggregate demand for the ith stock we use the same technique as employed by Lintner (1965a) with only one distinction: Lintner was allowed to sum up his equations for *all* investors. In our model, we are allowed to sum them up only for investors k who hold the security under consideration in their portfolios, since equation (4) (from which we derive equation (12)) includes the ith security only for investors k who hold it. After multiplying equation (12) by $T_k^2 \sigma_k^2$ and summing up *only for investors k who hold security i*, we obtain

$$(13) \quad [P_{i1} - P_{i0}(1 + r)] \sum_k T_k^2 \sigma_k^2 = \sum_k T_k(\mu_k - r) \left[N_{ik} \sigma_i^{*2} + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} N_{jk} \sigma_{ij}^* \right]$$

The equilibrium price of share i, P_{i0}, is given by

$$(14) \quad (1 + r)P_{i0} = P_{i1} - \left[\sum_k \left(T_k(\mu_k - r) \cdot \left[N_{ik} \sigma_i^{*2} + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} N_{jk} \sigma_{ij}^* \right] \right) \right] \div \sum_k T_k^2 \sigma_k^2$$

In order to derive a more comparable form for the equilibrium price as implied by the CAPM we multiply and divide by $[\Sigma_k T_k \cdot (\mu_k - r)]$ to obtain

$$(15) \quad (1 + r)P_{i0} = P_{i1} - \frac{\left[\sum_k T_k(\mu_k - r) \right]}{\sum_k T_k^2 \sigma_k^2} \cdot \frac{\sum_k \left(T_k(\mu_k - r) \left[N_{ik} \sigma_i^{*2} + \sum_{\substack{j=1 \\ j \neq i}}^{n_k} N_{jk} \sigma_{ij}^* \right] \right)}{\left[\sum_k T_k(\mu_k - r) \right]}$$

where P_{i0} is the equilibrium price of stock i as suggested by this model. The price of risk is given by $[\Sigma T_k(\mu_k - r)]/\Sigma T_k^2 \sigma_k^2$ and is relevant only for investors who hold security i. Obviously, investors who do not hold security i are faced by a different price of risk. Moreover, the same investor may face two (or more) different prices of risk, one appropriate for security i and one for security j. This may occur since the group of investors who hold security i is not nec-

essarily identical to the group of investors who hold security j. Thus, the term $\Sigma T_k \cdot (\mu_k - r)/\Sigma T_k^2\sigma_k^2$ (price of risk) is a function of the security under consideration, and is relevant only to investors who decide to hold this security in their portfolio.

The equilibrium formula given by equation (15) has very important implications for the empirical findings of the *CAPM*. To demonstrate, assume that all investors who hold security i hold also security j (namely, only two risky assets) and these investors purchase all the available securities of these two firms. For simplicity only, and without loss of generality, assume that $\mu_k - r$ is a constant (say $= A$) and that $T_k/\Sigma T_k = \alpha$ for all these investors. Thus (15) reduces to

$$(15')\quad (1 + r)P_{i0} = P_{i1} - \frac{\sum_k T_k(\mu_k - r)}{\sum_k T_k^2\sigma_k^2} \cdot \frac{\left[\sum_k T_k N_{ik}\sigma_i^{*2} + \sum_k T_k N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k}$$

On the basis of the above simplifying assumptions, we obtain from (15′)

$$(1 + r)P_{i0} = P_{i1} - \frac{\sum_k T_k(\mu_k - r)\alpha \sum_k\left[N_{ik}\sigma_i^{*2} + \sum_k N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k^2\sigma_k^2}$$

or

$$(15'')\quad (1 + r)P_{i0} = P_{i1} - \frac{\sum_k T_k(\mu_k - r)\alpha[N_i\sigma_i^{*2} + N_j\sigma_{ij}^*]}{\sum_k T_k^2\sigma_k^2}$$

since $\Sigma_k N_{ik} = N_i$, $\Sigma_k N_{jk} = N_j$, where N_i and N_j are the number of outstanding shares of i and j, respectively.

It can readily be seen from (15″) that the equilibrium price P_{i0} is a function of the ith security variance and of only *one* covariance, that is, its covariance with security j. Obviously, in such a case, we would expect that the ith security variance will play a central role in its equilibrium price determination, quite contrary to the result of the traditional *CAPM*. On the other hand, the traditional β_i (see equation (1)) has little to to with the determination P_{i0}, since β_i includes all the covariances (see equation (7)) while in the above example we have only one covariance. Note that few assumptions have been made in order to simplify the analysis. However, even when investors hold stocks of three or four companies, we still obtain the same result; the ith security variance is much more important in price determination than one would expect from the analysis of traditional *CAPM*. Empirical support to this theoretical result is given in Section IV.

For the specific case in which *all investors hold security* i, we sum up equation (12) for *all* investors k. Hence $\Sigma_k T_k(\mu_k - r)$ is the total aggregate excess dollar return of all investors' portfolios, which is equal to $T_0(\mu_m - r)$, where μ_m is the expected return on the market portfolio and $T_0 = \Sigma_k T_k$. However, $\Sigma_k T_k^2\sigma_k^2$ is not necessarily equal to $T_0^2\sigma_m^2$, and hence one does not have, even in the above specific case, the interpretation of the aggregate risk in the market as obtained when a perfect market is assumed. However, equation (15) can be written as

$$(1 + r)P_{i0} = P_{i1} - \frac{\left[\sum_k T_k(\mu_k - r)\right]}{T_0^2\sigma_m^2}\,\frac{T_0^2\sigma_m^2}{\sum_k T_k^2\sigma_k^2} \cdot \frac{\sum_k T_k(\mu_k - r)\left[N_{ik}\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n_k} N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k(\mu_k - r)}$$

If all investors hold security i, then $\Sigma_k T_k \cdot (\mu_k - r) = T_0(\mu_m - r)$ and the second term on the right-hand side is the market price of risk γ, when the *CAPM* is derived without constraint on the number of securities in the portfolio (see Lintner 1965a, p. 600).

Hence,

$$(16)\quad (1+r)P_{i0} = P_{i1} - \gamma\left(\frac{T_0^2\sigma_m^2}{\sum T_k^2\sigma_k^2}\right) \cdot \frac{\sum_k T_k(\mu_k - r)\left[N_{ik}\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n_k} N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k(\mu_k - r)}$$

or

$$(17)\quad (1+r)P_{i0} = P_{i1} - \gamma_1 \frac{\sum_k T_k(\mu_k - r)\left[N_{ik}\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n_k} N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k(\mu_k - r)}$$

where $$\gamma_1 = \frac{\gamma T_0^2\sigma_m^2}{\sum_k T_k^2\sigma_k^2}$$

Equation (17) is very similar to the classic relationship of the *CAPM* (see equation (20′)). The only two differences are: (a) now the securities' risk is given as the weighted average of the risks of each investor when the weights are $T_k(\mu_k - r)$, so that, the larger the investor's wealth (T_k), the greater his impact on price determination, and (b) the market price of risk γ_1 is defined somewhat differently from the well-known γ, as defined by Lintner (1965a). Thus, the classic *CAPM* may be the approximate equilibrium model for stocks of firms which are held by many investors (for example, AT&T), but not for small firms whose stocks are held by a relatively small group of investors.

If we relax the constraint that the kth investor holds only n_k securities, then each investor holds the market portfolio and hence[6] $\mu_k - r = \mu_m - r$, and $\sigma_m^2 = \sigma_k^2$, where μ_m and σ_m^2 are the expected rate of return and variance of the market portfolio, respectively.

[6]Recall that without loss of generality we deal only with the optimal unlevered portfolio The basic equilibrium equation (equation (6)) and hence all the other results derived from it are unchanged no matter if we deal with the levered or the unlevered portfolio See fn. 5.

On the basis of these assumptions we obtain the classic *CAPM* formula as a special case of the *GCAPM* suggested in this paper. In this case, equation (16) reduces to

$$(18)\quad (1+r)p_{i0} = p_{i1} - \gamma\frac{T_0^2}{\sum_k T_k^2} \cdot \frac{\sum_k T_k\left[N_{ik}\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n_k} N_{jk}\sigma_{ij}^*\right]}{\sum_k T_k}$$

But since the relaxation of the imperfection induces all investors to have the same investment strategy in risky assets, (see Sharpe and Lintner, 1965a) all of them hold all the risky assets $n_k = n$ and, also $N_{ik}/N_i = T_k/T_0$ and hence $N_{ik} = N_iT_k/T_0$ and $N_{jk} = N_jT_k/T_0$. By substituting the last reults in equation (18) we derive

$$(19)\quad (1+r)p_{i0} = p_{i1} - \gamma\frac{T_0^2}{\sum_k T_k^2}\;\frac{\sum_k T_k}{\sum_k T_k} \cdot \left[\frac{T_kN_i}{T_0}\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{T_kN_j}{T_0}\sigma_{ij}^*\right]$$

or

$$(20)\quad (1+r)p_{i0} = p_{i1} - \gamma\frac{T_0^2}{T_0\sum_k T_k^2} \cdot \frac{\sum_k T_k^2\left[N_i\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n} N_j\sigma_{ij}^*\right]}{\sum_k T_k}$$

Since $\Sigma_k T_k = T_0$, equation (20) reduces to the well-known equilibrium equation of the traditional *CAPM* (see Lintner 1965a, p. 600),

$$(20')\quad (1+r)p_{i0} = p_{i1} - \left[N_i\sigma_i^{*2} + \sum_{\substack{j=1\\ j\neq i}}^{n} N_j\sigma_{ij}^*\right]$$

Finally, I would like to emphasize the basic difference between equations (15) and (17). Equation (15), which I advocate, represents the most general form, and hence

only σ_{ij} of securities included in the kth investors' portfolios, are taken into account. However, if we assume unrealistically that security i is included in *all* investors' portfolios (equation (17)) then for an equilibrium price determination we must take into account the covariances σ_{ij} of all securities available in the market since we sum up in equation (17) for all k.

II. The Implication for the Empirical Findings

Recent empirical evidence indicates that the traditional *CAPM* does not explain the empirical data as well as might be expected. Douglas, using annual and quarterly data, shows that there is a significant relationship between the mean rate of return of a stock and its standard deviation—a fact which contradicts the *CAPM*. Lintner (1965b) regresses annual rates of return of 301 stocks over the period 1954 63. He estimates the systematic risk from time-series and then regresses the mean rate of return on the systematic risk and on the estimate of the residual variance (see equation (3)). His results, too, indicate that the theoretical model does not provide a satisfactory description of price behavior. Using annual data, Merton Miller and Myron Scholes confirm the basic results of Lintner and suggest possible explanations for the deviation between the model and the empirical evidence. Black, Jensen, and Scholes using monthly data also show that the model does not provide a satisfactory description of price behavior in the stock market.

In recent papers David Levhari and I have investigated the effect of the assumed investment horizon on the estimates of the systematic risk as well as on the other results implied by the *CAPM*. We have found that the investment horizon plays a crucial role in any econometric research and, particularly, in empirical work which tests the *CAPM*. However, in analyzing horizons ranging from one to twenty-four months, we have also found that the coefficient of the residual variance (γ_2 in equation (3)) remains significantly positive. In most cases, too, the residual variance explains price behavior even better than the estimates of the systematic risk (i.e., γ_1 in equation (3)). I demonstrate below that the fact that investors hold portfolios with only a few risky assets, rather than the market portfolio, provides a possible explanation for the three discrepancies between the theoretical model and the empirical findings obtained by various researchers.

Suppose that an investor holds a portfolio k whose random return is R_k, while the random return on the market portfolio is R_m. The expected return on R_k can be smaller or greater than the expected return of R_m. However, since R_k includes only a few securities while R_m consists of all securities available in the market, one would expect that the variance of R_m would be smaller than the variance of most selected portfolios, k. The relationship between R_k and R_m can be described as follows:

$$R_m = R_k + \psi \tag{21}$$

(alternatively, one can define this relationship in the form $R_m = a + bR_k + \psi$, see Miller and Scholes), where ψ is an error term. Let us now analyze the impact of the error in the variables given in (21), on empirical evidence related to the *CAPM*.

In the empirical research, the time-series regression is formulated as follows:

$$R_{it} = \alpha_i + \beta_i R_{mt} + e_t \tag{22}$$

where $\hat{\beta}_i$ derived from (22) is the estimate of the ith security systematic risk. Since the investors hold portfolio R_k rather than R_m, the true relationship is given by

$$R_{it} = \alpha^*_{ik} + \beta^*_{ik} R_{kt} + u_t \tag{23}$$

where β^*_{ik} is the kth investor's true systematic risk. We shall see that using (22) rather than (23) causes a certain bias in the estimate of the systematic risk. The estimate of $\hat{\beta}_i$ is given by

$$\hat{\beta}_i = \frac{cov(R_i, R_m)}{var(R_m)} = \frac{cov(R_i, R_k + \psi)}{var(R_k + \psi)} = \frac{cov(R_i, R_k) + cov(R_i, \psi)}{\sigma_k^2 + \sigma_\psi^2 + 2\,cov(R_k, \psi)} \tag{24}$$

If we divide by σ_k^2 and assume that the er-

rors are distributed independently of the true values (R_i and R_k), then the last term in the numerator, as well as the last term in the denominator, will tend to zero as the sample size increases indefinitely. Thus, $\hat{\beta}_i = cov(R_i, R_k)/(1 + \sigma_\Psi^2/\sigma_k^2)\sigma_k^2$. But since $cov(R_i, R_k)/\sigma_k^2 = \hat{\beta}^*_{ik}$ we finally obtain

$$\hat{\beta}_i = \frac{\beta^*_{ik}}{1 + \sigma_\Psi^2/\sigma_k^2} \tag{25}$$

Hence[7]

$$\hat{\beta}_i < \beta^*_{ik} \tag{26}$$

for all investors k, and hence $\hat{\beta}_i < \beta^*_i$ where β^*_i is a weighted average of β^*_{ik}. (I shall define this weighted average later on; see equation (35).)

Let us now investigate the impact of this bias in measuring the systematic risk, on the cross-section regression which is essential to an examination of the validity of the *CAPM* (see equation (1')). Since $\hat{\beta}_i$ is biased, one can write $\hat{\beta}_i$ as follows,

$$\hat{\beta}_i = \beta^*_i + \phi_i \tag{27}$$

where ϕ_i is an error term. Most empirical works carry out the cross-section regression in the following manner (see equation (1')):

$$\overline{R}_i - r = \gamma_0 + \gamma_1\hat{\beta}_i + e_i \tag{28}$$

while the true relationship is given by

$$\overline{R}_i - r = \gamma_0^* + \gamma_1^*\beta_i^* + e_i^* \tag{29}$$

where $\overline{R}_i$ is the average rate of return of the ith asset, r is the riskless interest rate, and $\hat{\beta}_i$ is the estimate of the systematic risk obtained from the time-series regression. Thus

$$\hat{\gamma}_1 = \frac{cov(\overline{R}_i, \hat{\beta}_i)}{\sigma^2(\hat{\beta}_i)} = \frac{cov(\overline{R}_i, \beta^*_i + \theta_i)}{\sigma^2(\beta^*_i) + \sigma^2(\theta_i) + 2cov(\beta^*_i, \theta_i)} = \frac{cov(\overline{R}_i, \beta^*_i) + cov(\overline{R}_i, \theta_i)}{\sigma^2(\beta^*_i) + \sigma^2(\theta_i) + 2cov(\beta^*_i, \theta_i)}$$

Dividing by $\sigma^2(\beta^*_i)$ and assuming that the error θ_i is uncorrelated with the values $\overline{R}_i$ and β^*_i, we obtain,

$$\hat{\gamma}_1 = \frac{cov(\overline{R}_i, \beta^*_i)/\sigma^2(\beta^*_i)}{1 + \sigma^2(\theta_i)/\sigma^2(\beta^*_i)}$$

or

$$\hat{\gamma}_1 = \frac{\hat{\gamma}_1^*}{1 + \sigma^2(\theta_i)/\sigma^2(\beta^*_i)}$$

and hence[8]

$$\hat{\gamma}_1 < \gamma_1^* \tag{30}$$

This may explain the result of most empirical studies where $\hat{\gamma}_1$ is below the value predicted by the *CAPM*

It has also been found in all empirical research that $\hat{\gamma}_0 > 0$, while, according to the *CAPM*, γ_0 should equal zero. This bias may be explained as follows: from equation (28) the estimate of γ_0 is given by

$$\hat{\gamma}_0 = \overline{\overline{R}} - \hat{\gamma}_1\overline{\hat{\beta}}$$

where $\overline{\overline{R}}$ is the average of the variables $\overline{R}_i - r$, and $\overline{\hat{\beta}}$ is the average of the estimates of the systematic risks $\hat{\beta}_i$ of all risky assets. However, the true relationship should be (from equation (29))

$$\hat{\gamma}_0 = \overline{\overline{R}} - \hat{\gamma}_1^*\hat{\beta}^*$$

since according to the above assumptions $\hat{\gamma}_1 < \hat{\gamma}_1^*$ and $\hat{\beta}_i < \hat{\beta}^*_i$, also $\hat{\gamma}_1\overline{\hat{\beta}} < \hat{\gamma}_1^*\overline{\hat{\beta}}^*$, hence we obtain the result $\hat{\gamma}_0 > \hat{\gamma}_0^* = 0$.

Apparently, the most disturbing empirical result is that $\hat{\gamma}_2$ (see equation (3)) is significantly greater than zero. The latter result, however, can be explained by the model presented in this paper. According to the *CAPM*, investors diversify in many securities, and hence, the residual variance $S^2_{e_i}$ should have no impact on the risk-return equilibrium relationship. The individual security's variance as well should have no impact on this relationship since the contribution of the individual risk is about $(1/n)\sigma^2(R_i)$ when n is the number of securities available in the market.[9] However, if

[7]In deriving (26) it is assumed that the errors are distributed independently of R_i and R_k. However, it is easy to verify that it is sufficient to require that u and Ψ are distributed independently and that the regression coefficient of R_k on Ψ_i is greater than -1

[8]Equation (30) is valid even under less restrictive assumptions (see fn 7)

[9]For simplicity's sake we assume that the investor diversify equally his resources among all securities. (See Miller and Scholes)

one assumes that investors hold undiversified portfolios which contain stocks of three or four companies (i.e., $n_k = 3, 4$) and that the ith security is not included in all portfolios, then the variance (and hence the residual variance) should have a strong impact on the risk-return relationship. Although we have already analyzed the role of the variance in price determination (see equation (15)), we can find a more transparent example by looking once again at equation (6). Rewriting (6) we obtain

$$(31) \qquad \mu_i - r = \frac{\mu_k - r}{\sigma_k^2} Cov(R_i, R_k)$$

Assuming, once again, for the sake of simplicity only, that the typical investor who holds security i will diversify equally between three stocks, we obtain

$$\mu_i - r = \frac{\mu_k - r}{\sigma_k^2} \cdot \left[Cov\left(R_i, \frac{1}{3} R_i + \frac{1}{3} R_{i-1} + \frac{1}{3} R_{i+1}\right)\right]$$

where i, $i-1$, and $i+1$ stand for the three securities included in the portfolio. Thus

$$(32) \qquad \mu_i - r = \frac{\mu_k - r}{\sigma_k^2}\left[\frac{1}{3}\sigma_{R_i}^2 + \frac{1}{3} Cov(R_i, R_{i-1}) + \frac{1}{3} Cov(R_i, R_{i+1})\right]$$

It is obvious from (32) that variance plays a central role in explaining the risk-return relationship. Moreover, one would expect that the individual variance would have greater impact on price determination than the β_i (as defined in equation (1)) since β_i has very little to do with the stock's risk when the portfolios include only a small number of different securities. Indeed, Douglas found that the coefficient of the variance is more important than the coefficient of the β in most periods covered in his empirical research.

To design a precise empirical study to test the model suggested in this paper is not an easy task since equation (6) includes a factor β_{ki} which varies from investor to investor. One has first to find a solution to the optimization problem with the constraint on the number of securities n_k, and also to know the amount invested by each investor in the stock market. To illustrate the difficulties involved in such an empirical test, let us reexamine equation (6′). When we multiply equation (6′) by T_k and sum up only for investors k who *hold security i*, we obtain

$$(33) \qquad \mu_i \sum_k T_k = r \sum_k T_k + \sum_k T_k(\mu_k - r)\beta_{ki}$$

or

$$(34) \qquad \mu_i = r + \sum_k T_k(\mu_k - r)\beta_{ki} / \sum_k T_k$$

By defining β_i^* as the weighted average, $\beta_i^* = \Sigma_k T_k(\mu_k - r)\beta_{ki}/\Sigma_k T_k(\mu_k - r)$ and $\Sigma_k T_k(\mu_k - r)/\Sigma_k T_k = \gamma_{1i}^*$ we can rewrite (34) as

$$(35) \qquad \mu_i = r + \gamma_{1i}^* \beta_i^*$$

where γ_{1i}^* varies from one security to another.

Equation (35) can then be used in order to test empirically the risk-return relationship as suggested in this paper. However, I would like to mention a few characteristic results as well as difficulties in testing this equation empirically: (a) Since $\hat{\beta}_i < \beta_{ik}$ for all k, $\hat{\beta}_i < \beta_i^*$ is also true, since β_i^* is a weighted average of β_{ik}. (b) $\gamma_{1i} = \Sigma T_k(\mu_k - r)/\Sigma T_k$ when we sum up only for investors k who hold security i. Thus, γ_{1i}^* varies from security to security, and any cross-section regression will provide an estimate of some average of all these γ_{1i}^*. (c) In order to test the *CAPM* in the present framework, one has to estimate first β_i^*, that is, to have information, not only on the selected portfolio by each investor k, but also on the relative size of his investment, $T_k/\Sigma T_k$. (d) Finally, it is worth mentioning that if all investors hold security i, $\gamma_{1i} = \Sigma_k T_k(\mu_k - r)/\Sigma_k T_k$, when we sum up for *all* investors k. Hence $\gamma_{1i} = \mu_m - r$, since in this case $\Sigma_k \cdot T_k = T_0$, and $\Sigma_k T_k(\mu_k - r) = T_0(\mu_m - r)$, where μ_m is the expected rate of return on the market portfolio.

Designing such an empirical research is beyond the scope of this paper. However, if the present form of the *CAPM* is correct,

TABLE 1—SECOND-PASS REGRESSION WITH MONTHLY DATA

$\bar{R}_i = \gamma_0$	+ $\gamma_1 \hat{\beta}_i$	+ $\gamma_2 \hat{S}^2_{e_i}$	+ $\gamma_3 \hat{\sigma}^2_i$	R^2
0.00894 (0.00096) $t = 9.3$	0.00196 (0.00094) $t = 2.1$			0.04
0.00985 (0.00057) $t = 17.3$			0.18369 (0.08956) $t = 2.0$	0.04
0.00999 (0.00053) $t = 19.0$		0.21916 (0.11129) $t = 2.0$		0.04
0.00914 (0.00099) $t = 9.2$	0.00117 (0.00136) $t = 0.86$		0.10404 (0.12865) $(t = 0.81)$	0.05
0.00899 (0.00096) $t = 9.3$	0.00136 (0.00110) $t = 1.2$	0.13736 (0.12909) $t = 1.1$		0.05

then, in spite of the fact that we do not have a perfect empirical procedure to test it, we expect the variance itself, σ_i^2, to provide a better explanation of price behavior than the traditional systematic risk, β_i.

III. The Empirical Findings

The monthly rates of return of a sample of 101 stocks traded on the New York Stock Exchange (NYSE) were calculated for the period 1948–68, that is, for each security there are 240 observations. Thus, if R_{i1}, $R_{i2}, \ldots, R_{i240}$ were the monthly rates of return, on the ith security, one can calculate the bimonthly rates of return, $R^*_{i1}, R^*_{i2}, \ldots, R^*_{i120}$ by substituting $(1 + R_{i1})(1 + R_{i2}) = 1 + R^*_{i1}$, $(1 + R_{i3})(1 + R_{i4}) = 1 + R^*_{i2}$ etc., where R^*_i $(i = 1, 2, \ldots, 120)$ are the rates of return for an investment horizon of two months. Note that, by using a horizon of two months, we subdivided the period 1948–68 to 120 time units rather than to 240 time units, without changing the length of the period covered by the empirical research: namely, twenty years. Similarly, if we had used annual rates of return, we would have only 20 observations. As a proxy to the market portfolios I used the Fisher Arithmetic Index, which assumes an equal investment in each of the NYSE stocks.

In this paper we examine the following linear regressions, with monthly data, semi-annual data and annual data:

$$\bar{R}_i - r = f(\hat{\beta}_i)$$
$$\bar{R}_i - r = f(\hat{S}^2_{e_i})$$
$$\bar{R}_i - r = f(\hat{\sigma}^2_i)$$
$$\bar{R}_i - r = f(\hat{\beta}_i, \hat{S}^2_{e_i})$$
$$\bar{R}_i - r = f(\hat{\beta}_i, \hat{\sigma}^2_i)$$

where $\bar{R}_i$ is the average rate of return on the ith security, r is the rate of return on riskless assets,[10] and $\hat{\beta}_i$ is the systematic risk estimated from the time-series regressions; $\hat{S}^2_{e_i}$ is the residual variance (taken also from the time-series regressions) and $\hat{\sigma}^2_i$ stands for the estimate of the ith security variance.

These regressions are run for three different investment horizons (one, six, and twelve months) since it has been shown that

[10]The rates of return on Treasury Bills as well as on government bonds were taken from various issues of the *Federal Reserve Bulletin*. The sample of shares was taken from the return file of the CRSP tape. Note that in estimating *Beta*, and in the cross-section regression we employ the same set of data. This may cause some statistical bias. However, I believe that by a division of the period to two superiods (one for estimating *Beta* and the other for the cross-section regression) one may lose many observations, which is undesirable. Moreover, the *Beta* may change from period to period which decreases the reliability of this procedure.

TABLE 2—SECOND-PASS REGRESSION WITH SEMIANNUAL DATA

$\bar{R}_i = \gamma_0$	+	$\gamma_1\hat{\beta}_1$	+	$\gamma_2\hat{S}^2_{e_i}$	+	$\gamma_3\hat{\sigma}^2_i$	R^2
0 0493 (0 0048) t = 10 2		0 0219 (0 0046) t = 4 7					0.19
0.0583 (0.0030) t = 19 4						0 2630 0 0517 t = 5 1	0 21
0 0603 (0 0029) t = 20.8				0 3378 (0 0747) t = 4 5			0.17
0.0528 (0 0050) t = 10 6		0 0099 (0 0072) t = 1 4				0 1771 (0 0808) t = 2.2	0 23
0 0494 (0 0047) t = 10 6		0 0151 (0.0052) t = 2 9		0.2164 (0.0834) t = 2 6			0 24

in the cross-section regression, the estimate of the systematic risk and the other parameters, (for example, R^2) are very sensitive to the assumed investment horizon (see Levhari and the author). Tables 1, 2, and 3 summarize the empirical findings for one-month, six-month, and twelve-month horizons, respectively In Table 1, most of the regression coefficients are insignificant, and the coefficient of correlation is very low (less than 5 percent), indicating that the assumption of a one-month horizon is a very poor assumption.[11]

Moving to Table 2, we still obtain a low R^2. However, even from this table one can see that the simple regression $\bar{R}_i - r = f(\hat{\sigma}^2_i)$ yields a better (or at least not a worse) explanation than the regression $\bar{R}_i - r = f(\hat{\beta}_i)$, which is implied by the *CAPM*. Using the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{S}^2_{e_i})$ we find that the coefficient of $\hat{\beta}_i$ as well as the coefficient of $\hat{S}^2_{e_i}$ are statistically significant. The fact that the coefficient of $\hat{S}^2_{e_i}$ is significant is quite obvious from the above analysis. It is particularly obvious from the fact that each investor holds only a few securities in his portfolio, since $\hat{S}^2_{e_i}$ serves as a proxy to $\hat{\sigma}^2_i$. Indeed the R^2 between $\hat{S}^2_{e_i}$ and $\hat{\sigma}^2_i$ in this sample is 0.80. The fact that the coefficient of $\hat{\beta}_i$ is significant can be explained by the fact that the estimate of $\hat{\beta}_i$ is also correlated with $\hat{\sigma}^2_i$. Thus, even though β_i has little to do with the security risk, the regression coefficient of β_i is positive, since β_i is positively correlated with a main component of the true risk (σ^2_i). (Indeed, as we shall see below, β_i plays no role in price determination.) Such seems to be the case for the present sample where the relation between $\hat{\beta}_i$ and $\hat{\sigma}^2_i$ is[12]

$$\hat{\beta}_i = \underset{(0.06)}{0.68} + \underset{(0.32)}{2\ 78}\,\hat{\sigma}^2_i$$

$$t = \quad 11.4 \qquad 8.63 \qquad\qquad R^2 = 0.43$$

When we run the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{\sigma}^2_i)$ we find that the coefficient of $\hat{\sigma}^2_i$ is positive and significant, while the coefficient of $\hat{\beta}_i$ becomes insignificant. Once again, the simple model $\bar{R}_i - r = f(\hat{\sigma}^2_i)$ can explain price behavior almost as well as any other suggested model.

[11]One can easily increase the R^2 by running rates of a group of securities $\bar{R}_i$ on the market portfolio However, since the *CAPM* should hold for individual securities, I think that a high R^2 which is achieved by grouping neither confirms nor refutes the *CAPM*

[12]Miller and Scholes have found that the estimates of the systematic risk β_i is also correlated with the residual variance They found in their sample R^2 = 0 17 while similar regression of the present sample yields R^2 = 0 14

TABLE 3—SECOND-PASS REGRESSION WITH ANNUAL DATA

$\bar{R}_i = \gamma_0$	+	$\gamma_1\hat{\beta}_1$	+	$\gamma_2\hat{S}^2_{e_i}$	+	$\gamma_3\hat{\sigma}^2_i$	R^2
0.109 (0.009) $t = 12.0$		0.037 (0.008) $t = 5.1$					0.21
0.122 (0.005) $t = 22.9$						0.219 (0.029) $t = 7.7$	0.38
0.126 (0.005) $t = 23.4$				0.248 (0.036) $t = 6.8$			0.32
0.117 (0.008) $t = 14.2$		0.008 (0.009) $t = 0.9$				0.197 (0.038) $t = 5.2$	0.38
0.106 (0.008) $t = 13.2$		0.024 (0.007) $t = 3.3$		0.201 (0.038) $t = 5.3$			0.39

Table 3 deals with annual data and confirms the previous results. First, we note that our results are very similar to those obtained by Miller and Scholes (who also used annual data) in spite of the fact that a different sample of data is used. We find the R^2 of the regression $\bar{R}_i - r = f(\beta_i)$ to be 21 percent in comparison to 19 percent in their research; for the regression $\bar{R}_i - r = f(\hat{S}^2_{e_i})$ we obtain 32 percent in comparison to their 28 percent; and finally, for the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{S}^2_{e_i})$ we find R^2 to be equal to 39 percent in comparison to 34 percent that they obtain. With annual data, all the regression coefficients are positive and significant in my research as well as in Miller and Scholes' research. However, in Table 3, I present two more regressions which do not appear in Miller and Scholes' paper. These two regressions confirm the previous results of the semiannual data which can be summarized as follows: (a) The simple regression $\bar{R}_i - r = f(\sigma^2_i)$ yields R^2 of 38 percent. This is only 1 percent less than the more complicated regression $\bar{R}_i = f(\hat{\beta}_i, \hat{S}^2_{e_i})$ which has been employed in most empirical studies that test the validity of the *CAPM*. (b) When we run the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{\sigma}^2_i)$ rather than $\bar{R}_i - r = f(\hat{\beta}_i, \hat{S}^2_{e_i})$, we find that the conventional estimate of the systematic risk $\hat{\beta}_i$ adds nothing to the explanation of price behavior. The coefficient of the systematic risk is very small and statistically insignificant (t value = 0.9). (c) If one had to choose between the traditional *CAPM* (i.e., $\bar{R}_i - r = f(\hat{\beta}_i)$ and the simple model $\bar{R}_i - r = f(\hat{\sigma}^2_i)$, one would note that the latter performs much better, with $R^2 = 38$ percent compared to only $R^2 = 21$ percent for the previous model.

IV. Concluding Remarks

The assumption of the perfect indivisibility of an investment and of the absence of transaction costs in the stock market, induces a theoretical result which asserts that each investor holds in his portfolio all the securities available in the market. It is obvious that the above assumption does not conform to reality, since many investors hold stocks of only one company, and most individuals hold stocks of less than four companies. Nor can we accept *CAPM* on a positive ground since it performs quite poorly in explaining price behavior.

In this paper, I have relaxed the assumption of a perfect market, and hence, the kth investor holds stocks of n_k companies in his portfolio where n_k can be very small (i.e., 1, 2, etc.). We first derive an equilibrium re-

lationship between the return and risk of each security. We have found that the well-known systematic risk of the traditional *CAPM*, β_i, has little to do with equilibrium price determination. On the other hand, β_i^*, which is a weighted average of the kth investor systematic risk β_{ik}, is the correct measure of the ith security risk. Since σ_i^2 is a major component of β_{ik}, it plays a crucial role in the risk measure of each stock, quite contrary to the equilibrium results of the capital asset pricing model. When we impose the assumption of a perfect market and assume that investors hold all the available risky assets, (i.e., $n_k = n$), we obtain the well-known form of the *CAPM* as a special case of the *GCAPM* developed in this paper. The suggested model developed here, based on the fact that individuals hold relatively undiversified portfolios, explains the empirical results of the cross-section regression which have been found in most empirical studies.

The empirical findings support the theoretical results. The simple regression $\bar{R}_i - r = f(\hat{\sigma}_i^2)$ performs much better than the regression $\bar{R}_i - r = f(\hat{\beta}_i)$. The fact that $\bar{R}_i$ and β_i are positively correlated is caused simply by the fact that β_i and σ_i^2 are positively correlated, and that β_i serves as a proxy to the true risk component σ_i^2. In the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{S}_{e_i}^2)$, the coefficient of $\hat{\beta}_i$ as well as of $\hat{S}_{e_i}^2$ is positive and significant. The latter results have been found in other studies as well as in this paper. However, we claim that the coefficients of $\hat{\beta}_i$ and $\hat{S}_{e_i}^2$ are upward biased since $\hat{\beta}_i$ as well as $\hat{S}_{e_i}^2$ are positively correlated with $\hat{\sigma}_i^2$. Indeed, when we ran the regression $\bar{R}_i - r = f(\hat{\beta}_i, \hat{\sigma}_i^2)$ we found that the regression coefficient of $\hat{\sigma}_i^2$ was significant whereas the coefficient of $\hat{\beta}_i$ did not differ significantly from zero. This confirms the notion that, in an imperfect market, $\hat{\beta}_i$ plays no role, or at least a negligible role in price determination.

I would like to mention that σ_i^2 plays a central role in the risk-return relationship, but it is not the only measure of the ith security risk. The variance is only one component in this risk, and an empirical test should be designed in order to examine the validity of the *CAPM* in its imperfect form. Designing a precise empirical test which examines the validity of the *CAPM* in an imperfect market, is not an easy task, and is beyond the scope of this paper.

Finally, I think that the true risk index of the ith security is determined in the market, somewhere between the ith variance $\hat{\sigma}_i$, and the more sophisticated index β_i, as implied by the *CAPM*. For securities which are widely held (i.e., AT&T) we expect that *Beta* will provide a better explanation for price behavior[13] (see equation (17)), while for most securities, which are not held by many investors we would expect that the variance σ_i^2 would provide a better explanation for price behavior.

[13]Blume and Friend (1974) who tested the *Beta* and another quality rating index as measures of risk come to the conclusion that the *Beta* index performs relatively better for stocks with large market values. On the assumption that large market values implies also that the stocks are held by relatively many investors, this finding is consistent with my theoretical argument.

REFERENCES

F. Black, M. C. Jensen, and M. Scholes, "The Capital Asset Model: Some Empirical Tests," in Michael C. Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1972.

M. E. Blume and I. Friend, "Risk, Investment Strategy and the Long-Run Rates of Return," *Rev. Econ. Statist.*, Aug. 1974, *56*, 259–69.

______ and ______, "The Asset Structure of Individual Portfolios and Some Implications for Utility Functions," *J. Finance*, May 1975, *30*, 585–603.

______, J. Crockett, and I. Friend, "Stock Ownership in the United States: Characteristics and Trends," *Surv. Curr. Bus.*, Nov. 1974, *54*, 16–40.

G. W. Douglas, "Risk in the Equity Markets: An Empirical Appraisal of Market Efficiency," *Yale Econ. Essays*, Spring 1969, *9*, 3–45.

I. Friend and M. Blume, "Measurement of Portfolio Performance Under Uncertainty," *Amer. Econ. Rev.*, Sept. 1970, *60*,

561 75.

D. Levhari and H. Levy, "The Capital Asset Pricing Model and the Investment Horizon," *Rev. Econ. Statist.*, Feb. 1977, *59*, 92 104.

H. Levy, "Portfolio Performance and the Investment Horizon," *Manage. Sci.*, Aug. 1972, *36*, 645 53.

J. Lintner, (1965a) "Security Prices, Risk, and Maximal Gains from Diversification," *J Finance*, Dec. 1965, *20*, 587 616.

———, (1965b) "Security Prices and Risk: The Theory of Comparative Analysis of AT&T and Leading Industrials," paper presented at the Conference on the Economics of Regulated Public Utilities, Chicago, June 1965

———, "The Aggregation of Investors' Diverse Judgements and Preferences in Purely Competitive Security Markets," *J. Finance Quant Anal*, Dec. 1969, *4*, 347 400.

Harry M. Markowitz, "Portfolio Selection," *J. Finance*, Mar. 1952, *6*, 77-91.

———, *Portfolio Selection: Efficient Diversification of Investment*, New York 1959.

M. Miller and M. Scholes, "Rate of Return in Relation to Risk: A Reexamination of Some Recent Findings," in Michael C. Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1972.

W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425-42.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *26*, 65 86.

J. L. Treynor, "Toward a Theory of Market Value of Risky Assets," unpublished paper, 1961.

Board of Governors of the Federal Reserve System, *Fed Res Bull.*, Washington, various issues

On the Almost Total Inadequacy of Keynesian Balance-of-Payments Theory

By EDWARD A. KUSKA*

The object of this paper is, unfortunately, destructive. It is to argue that almost all of the models in the Keynesian balance-of-payments literature suffer from internal contradictions and deficiencies which make them unsuitable for balance-of-payments theory. Doubling the Keynesian closed-economy equations does not provide an appropriate description of the international economy. More careful attention needs to be given to portraying the money markets than is usually the case, and assets must be included in the various demand functions in any acceptable analysis.

Contributions employing the monetary approach to balance-of-payments theory are largely free of the criticisms made in this paper; see, for example, those by Frank Hahn, Murray C. Kemp (1962, 1964, 1970), Takashi Negishi, Rudiger Dornbusch (1973a,b), Pentti J. K. Kouri and Michael G. Porter, Michael Mussa, Jacob A. Frenkel and Carlos A. Rodriguez, Richard K. Anderson and Akira Takayama, and the author (1970, 1972, 1975, 1976, 1977).

The fact that theorists using a different approach have written papers which do not fall into the difficulties discussed below does not, however, necessarily imply either that the profession at large is, or the particular theorists involved were, aware of these problems in the Keynesian literature. We have, for instance, the following quotation from a well-known article in the monetarist literature by Dornbusch.

> Furthermore, to the extent that the monetary approach does suggest that monetary changes do have short-run effects on real variables, it becomes an empirical issue to determine whether observed behavior is compatible with that theory or more readily agrees with the "Keynesian" variants of complete specialization and money prices that are fixed in terms of the producer's currency.
>
> While both the Keynesian approach to devaluation and the monetary formulation of that problem are theoretically correct and can only be questioned with respect to the attractiveness of the underlying assumptions, this is not true of a large body of literature that discusses the effects of a devaluation in the framework of the standard barter model of trade. [1973a, p. 894]

A second example is a paper by Alexander Swoboda and Dornbusch which considers the long-run behavior of several Keynesian models, each of which is involved in at least one of the contradictions and shortcomings stated in Propositions 1 through 4 of this paper. (Their article is discussed in Section V.)

In addition there is John F. Kyle's, *The Balance of Payments in a Monetary Economy*, a revision of his doctoral dissertation which won the prestigious Irving Fisher Award in 1973. Kyle's chapter 5 discusses the monetary approach models without serious error. In his earlier chapters, however, in which he surveys the Keynesian literature, he seems unaware of the criticisms levied here; and, indeed, his own attempt to provide a reconciliation between the elasticities and absorption approaches is subject to the strictures presented in Propositions 1, 2, and 3 in Section IV below.

In fairness, however, it must be noted that, although the magnitude of the criticisms which may be leveled against the Keynesian models has not been fully ap-

*Lecturer, London School of Economics

preciated, there does seem to have developed a widespread feeling that these models do omit essential components of the analysis of the balance of payments. And, indeed, several writers whose earlier articles are criticized in this paper have more recently written contributions along these lines.

To economize on space, I shall throughout the body of this paper discuss the various models under the assumption of fixed exchange rates. The minor alterations to the exposition which are necessary when the international monetary regime is one of flexible rates are left to the reader.

I. The Closed Economy Hicksian *IS-LM* Model

Consider first the *IS-LM* model. We may write it as

$$(1)\qquad Y = C + I$$

$$(2)\qquad C = C(Y,r)$$

$$(3)\qquad I = I(r)$$

$$(4)\qquad M = L(Y,r)$$

where Y is income, C is consumption, I is investment, r is the rate of interest, M is the supply of money, and the price level is assumed to be constant and set equal to unity.

Equation (1) is the condition for equilibrium in the goods market, (4) is the equilibrium condition for the money market, and Walras' Law permits the suppression of the equilibrium condition for the remaining good, bonds. The aggregate budget constraint for the model may be written as

$$(5)\quad [C + I - Y] + \frac{1}{r}[H(\ ) - A] + [L(Y,r) - M] = 0$$

where A is the quantity, $1/r$ is the price, and $H(\ )$ is the demand function for bonds; and where, for the moment, we have left the arguments of that function unspecified.

The excess demand function for bonds may be expressed in two different ways, i.e., either as

$$(6)\qquad E_A = H(\ ) - A$$

or, because of the budget constraint and equations (2) and (3), as

$$(7)\quad E_A = r[M - L(Y,r) + Y - C(Y,r) - I(r)]$$

If these two equations are not to be inconsistent as they stand, $H(\ )$ must take the following form

$$(8)\qquad H(\ ) = \overline{H}(Y,r) + A + rM$$

where $\overline{H}(Y,r)$ is an expression in Y and r. The assumptions therefore embodied in this model are economically most implausible. Firstly, *any* exogenous increase in the quantity of bonds will be willingly held no matter what the values of the level of income, the rate of interest, or the quantity of money. This allows the total wealth of the private sector to become indefinitely large without influencing any other variable in the model except the demand for bonds—if the increase comes solely through increments in the stock of that good.[1] Secondly, any increase in the money stock is, for given levels of income, the stock of bonds, and the rate of interest, entirely used to purchase bonds. None is held as additional cash balances or spent on consumption. (As an indication of the sort of implication which may be derived from this type of model, it may be noted that it makes no difference here whether the supply of money is increased through cash grants or through open market operations.)

Now, while leaving assets out of the demand functions of the nonsuppressed equations of a closed model is obviously unrealistic, it does not necessarily involve the model in inconsistency. In open economies, however, the omission is crucially unrealistic, and in a number of instances in the literature it has led authors into one of the mathematical contradictions which has induced the writing of this paper. This point

[1] There is, of course, the argument that when expected future tax liabilities are considered, an increase in government bonds does not increase the private sector's wealth. This argument does not imply, however, that the supply of bonds has no effect whatsoever on the other markets of the system.

is however more easily discussed in Section III below following the presentation of the model embodied in equations (37)-(47) and (48)-(49).

II. The Two-Economy *IS-LM* Model

Consider now a fixed-exchange rate, *IS-LM* two-economy world, and suppose *for simplicity only* that the capital markets are completely segregated. The equations in the literature are then often written as some variant of the following:

$$(9) \qquad Y = C + I + X - eX^*$$

$$(10) \qquad Y^* = C^* + I^* + X^* - \frac{1}{e} X$$

$$(11) \qquad M = L(Y,r)$$

$$(12) \qquad M^* = L^*(Y^*,r^*)$$

$$(13) \qquad C = C(Y,r)$$

$$(14) \qquad C^* = C^*(Y^*,r^*)$$

$$(15) \qquad I = I(r)$$

$$(16) \qquad I^* = I^*(r^*)$$

$$(17) \qquad X = X\left(Y^*, \frac{1}{e}\right)$$

$$(18) \qquad X^* = X^*(Y,e)$$

$$(19) \qquad B_T = X - eX^*$$

$$(20) \qquad B_T^* = X^* - \frac{1}{e} X$$

where unstarred variables refer to the home economy and starred variables to the foreign economy. The variable Y, C, I, r and M are as defined for equations (1) and (2), X is exports, B_T is the balance of trade, and e is the exchange rate, defined as the home currency price of foreign money. The exogenous variables are usually e, M, and M^*, although sometimes r and r^* replace the latter two.

There are three criticisms to be made here. The first is the same as that made in the previous section, that is, that the model requires rather peculiar excess demand functions for bonds, if each country's budget constraint is not to be violated. The second criticism, however, is much more crucial since it involves a basic contradiction in the model. It is this. By requiring in (11) and (12) equality between the demands and supplies of money, *the model forces the balance of payments of each country to be zero.* It then becomes nonsensical to investigate the effects on either economy's balance of payments of variations in the exogenous variables.

The argument to demonstrate this proposition runs as follows. First, let M, M^*, and e be the exogenous variables. The aggregate budget constraints for the home and foreign economies are then

$$(21) \qquad [Y - C - I + eX^*] + [-eX^*] + \frac{1}{r}[A - H(\ \ )] + [M - L(Y,r)] = 0$$

$$(22) \qquad \left[Y^* - C^* - I^* + \frac{1}{e} X\right] + \left[-\frac{1}{e} X\right] + \frac{1}{r^*}[A^* - H^*(\ \ )] + [M^* - L^*(Y^*,r^*)] = 0$$

where the excess supply of each good is bracketed and, as before, A and A^* and $H(\ \ )$ and $H^*(\ \ )$ are the supplies and demands for bonds. Since the literature provides little guidance on the matter, I again leave the arguments in the demand functions for bonds unspecified. The overall balance of payments in an economy's own currency, B and B^*, may be defined, both *ex ante* and *ex post*, as the value of the excess supplies of all other goods except money, i.e., as

$$(23) \qquad B = [Y - C - I + eX^*] + [-eX^*] + \frac{1}{r}[A - H(\ \ )]$$

$$(24) \qquad B^* = \left[Y^* - C^* - I^* + \frac{1}{e} X\right] + \left[-\frac{1}{e} X\right] + \frac{1}{r^*}[A^* - H^*(\ \ )]$$

Equations (21)-(22) and (23)-(24) together imply

$$(25) \qquad B = L(Y,r) - M$$

$$(26) \qquad B^* = L^*(Y^*,r^*) - M^*$$

In addition in equilibrium, if each bond

market clears, we have from equations (9)-(20), (23)-(24), and (25)-(26),

$$(27)\quad B = B_T = X - eX^* = L(Y,r) - M = 0$$

$$(28)\quad B^* = B_T^* = X^* - \frac{1}{e}X = L^*(Y^*,r^*) - M^* = 0$$

The model outlined by equations (9)-(20) with the money supplies exogenous cannot therefore be seriously employed for balance-of-payments analysis. This, as detailed in Section V, effectively rejects a large proportion of the Keynesian balance-of-payments literature.

Suppose, on the other hand, that M and M^* are treated as endogenous with r and r^* exogenous. The economic sense of this presumably is that the monetary authorities step in to buy or sell bonds to keep the rates of interest constant. Rewrite the private sector budget constraints as

$$(29)\quad [Y - C - I + eX^*] + [-eX^*] + \frac{1}{r}[A - H_P(\;\;)] + [M_P - L(Y,r)] = 0$$

$$(30)\quad \left[Y^* - C^* - I^* + \frac{1}{e}X\right] + \left[-\frac{1}{e}X\right] + \frac{1}{r^*}[A^* - H_P^*(\;\;)] + [M_P^* - L^*(Y^*,r^*)] = 0$$

where the only differences between these two equations and (21) and (22) are that the initial money stocks of the private sectors and their demands for bonds now have a subscript P as a distinguishing mark.

The monetary authorities' budget constraints are

$$(31)\quad \frac{1}{r}[-H_G] + [\Delta M_G] = 0$$

$$(32)\quad \frac{1}{r^*}[-H_G^*] + [\Delta M_G^*] = 0$$

where H_G and H_G^* are the monetary authorities' demands for bonds and ΔM_G and ΔM_G^* are the increments in the money stocks brought about by the authorities' activities in the bond markets.

The aggregate budget constraints for each country are obtained by summing equations (29)-(30) and (31)-(32) country by country. This operation yields

$$(33)\quad [Y - C - I + eX^*] + [-eX^*] + \frac{1}{r}[A - H_P(\;\;) - H_G] + [M - L(Y,r)] = 0$$

$$(34)\quad \left[Y^* - C^* - I^* + \frac{1}{e}X\right] + \left[-\frac{1}{e}X\right] + \frac{1}{r^*}[A^* - H_P^*(\;\;) - H_G^*] + [M^* - L^*(Y^*,r^*)] = 0$$

where now M and M^*, the total supplies of money, are given by

$$(35)\quad M = M_P + \Delta M_G$$

$$(36)\quad M^* = M_P^* + \Delta M_G^*$$

Equations (33)-(34) and (23)-(24) again imply equations (25)-(26) and, if the bond markets clear, equations (27)-(28). Therefore, in the endogenous money supply case as well, the inclusion in the model of equations such as (11) and (12), which clear the two money markets, forces each country's balance of payments to vanish.

The third criticism of the model results from the fact that both bond market-clearing equations are omitted in (9)-(20). Walras' Law allows for the suppression of one equation, but nothing provides for the omission of the second, and without it the model does not require equilibrium in either of these markets.[2] This fact alone should have been indication enough to theorists employing variations of this model that something was wrong.

III. A Restated Two-Economy Model

The difficulties with the model presented in equations (9)-(20), aside from the pe-

[2]Therefore the move from equations (25)-(26) to equations (27)-(28) was not really justified on mathematical grounds. This does not alter the conclusion presented there, however. In a properly specified model, external equilibrium is implied by the requirement of equality between the demand and supply of money in each of the two countries.

culiarities of the excess demand functions for bonds, are associated with the descriptions of the money markets and the omission of the bond market-clearing equations. Suppose, to deal with the latter problem, we borrow a leaf from Robert Mundell (1963, 1964) and assume perfect capital mobility, so that we need include only one bond market and one rate of interest. In this case it is perfectly valid to use Walras' Law to suppress the market-clearing equation for that good. For the former problem we may follow a number of writers who have altered the money market equations to allow currency flows between the two economies so that balance-of-payments disequilibria are possible.

A model which includes these adjustments is the following:

$$(37)\quad Y = C + I + X - eX^*$$

$$(38)\quad Y^* = C^* + I^* + X^* - \frac{1}{e}X$$

$$(39)\quad M + eM^* - L(Y,r) - eL^*(Y^*,r) = 0$$

$$(40)\quad C = C(Y,r)$$

$$(41)\quad C^* = C^*(Y^*,r)$$

$$(42)\quad I = I(r)$$

$$(43)\quad I^* = I^*(r)$$

$$(44)\quad X = X\left(Y^*,\frac{1}{e}\right)$$

$$(45)\quad X^* = X^*(Y,e)$$

$$(46)\quad B = L(Y,r) - M$$

$$(47)\quad B^* = L^*(Y^*,r) - M^*$$

where r is now the world interest rate and all other variables are as defined in the previous section. We continue to assume that equations (21)–(22) and (23)-(24) hold (where now $r = r^*$) so that the balance of payments may be expressed as in (25)–(26) or, as we have done here, as in (46) and (47).

It is the operations of the exchange authorities in making the two moneys perfect substitutes which bring about a single market for that good. The terms M and M^* denote the initial supplies of money in each period, but of course these alter over time according to the currency flows given by (46) and (47). Therefore if there is no domestic credit creation or destruction we have

$$(48)\quad M_t = L(Y_{t-1},r_{t-1})$$

$$(49)\quad M_t^* = L^*(Y_{t-1}^*,r_{t-1})$$

where the subscript t denotes period t.[3]

Is this model, except for the specification of the excess-demand function for bonds, acceptable as a description of the international economy? The answer is quite simply "No." It still remains entirely unsuitable.

Suppose for the moment that equations (37)–(47), (21)–(22), (23)–(24), and (48)–(49) are valid and that an equilibrium exists. Now let the home economy devalue by increasing e, and consider the effect of this policy on the balance of payments. Theorists in the Keynesian tradition usually attacked this problem by deriving conditions for the sign of the total derivative of the balance of trade, $B_T = X - (1/e)X^*$, to be positive, although a number of contributions did include the capital account to derive the effect on the overall balance of payments. However, because of equations (21)–(22), (23)–(24), and (25)-(26), it is possible to determine the effect of devaluation on the overall balance of payments by investigating its effect on the excess demand for money as given by (46) and (47).

Now, in the first postdevaluation equilibrium, equation (39) sets the total world demand for money equal to its supply, when these variables are expressed in a common currency; and (46) and (47) yield the amount of this good which is required by the equilibrium to pass over the exchanges.

Consider the second postdevaluation equilibrium, given by the same equations, where now, however, because of equations (48)–(49), the initial money balance of each economy is equal to its demand in the previous period. The equilibrium values of all dependent variables in the equations except B and B^* are given by equations (37)-(45) and these equations depend only on the total quantity of money, $M + eM^*$, not its distribution between the two countries.

[3] The introduction of fractional reserve banking systems would alter none of the essentials of the argument which follows.

Therefore, the equilibrium values of the dependent variables in these equations are the same in the second postdevaluation equilibrium as they were in the first. This is true, in particular, of Y, Y^*, and r, and therefore of $L(Y,r)$ and $L^*(Y^*,r)$.

Equations (48) (49), (46) and (47) allow B and B^* in the second postdevaluation period to be written as

$$(50) \quad B_{t_0+1} = L(Y_{t_0+1}, r_{t_0+1}) - L(Y_{t_0}, r_{t_0})$$

$$(51) \quad B^*_{t_0+1} = L^*(Y^*_{t_0+1}, r_{t_0+1}) - L^*(Y^*_{t_0}, r_{t_0})$$

where t_0 denotes the first postdevaluation period and $t_0 + 1$ the second. From the above argument, however, we know that $Y_{t_0+1} = Y_{t_0}$, $Y^*_{t_0+1} = Y^*_{t_0}$, and $r_{t_0+1} = r_{t_0}$, so that *each country's overall balance of payments vanishes in the second postdevaluation period.* And of course this will be true of every succeeding period as well, unless one of the parameters of the model is altered.

The disturbance mentioned here was a change in the exchange rate, but the same result would follow a variation in any parameter. In this model the required currency flows resulting from any disturbance take place completely in the period in which that disturbance occurs. Therefore each country in the model reexhibits balance-of-payments equilibrium in the second and every succeeding period following such a disturbance. A model which so flies in the face of even the most casual observation of the real world is not in any way acceptable for balance-of-payments analysis. And no additional assumptions of sterilization or similar behavior by the monetary authorities can disguise this basic inadequacy.

IV. Summary

We now have the following propositions:

PROPOSITION 1: *Models which do not include assets in the demand functions of their nonsuppressed equations require the demand function in the suppressed (because of Walras' Law) equation to be peculiarly and asymmetrically specified.*

PROPOSITION 2: *Models which include equations which bring each country's demand and supply of money into equality have zero overall balance-of-payments figures in all periods. This is true whether the supplies of money are taken to be endogenous or exogenous. If, in addition, the model includes other equations which allow the overall balance of payments to be nonzero, it is contradictory*

PROPOSITION 3: *Models in which more than one market-clearing equation is suppressed do not require equilibrium in any of the excluded markets. (In the literature, this usually happens with the bond markets.)*

PROPOSITION 4: *Consider models* (a) *which do not include assets (in particular, money) in the demand functions of their nonsuppressed equations and* (b) *which allow intercountry flows of currency by essentially setting the world demand for money equal to its supply. These models have zero balance-of-payments values in the second and every succeeding postdisturbance period. And, if they also include balance-of-payments functions which are not so constrained, they too are contradictory.*

V. Illustrations from the Literature

Lest the reader feels that the foregoing is all too obvious, we may consider a selection of papers from the literature. I make no attempt at comprehensiveness; instead I intend to illustrate how often inconsistent models have been presented in what continue to be quite highly regarded papers. (I might add, however, that it is extremely difficult to find models in the Keynesian tradition which do not fall into inconsistency.)

We shall first consider James Meade's contribution. Meade presents his complete model in Section III (pp. 10-14). His equations (1.21) and (1.22) are the demand functions for money in country A and country B, respectively, and his (1.23) equates the related demands for gold to the world's supply. These three equations reduce essentially to (39) in my Section III above.

Meade first considers the fixed exchange rate case under the heading, "The Gold Standard," on page 13. Here he sets country A's overall balance of payments, given by his equation (1.20), equal to zero, "since the system is in external equilibrium." It is not at all obvious why Meade felt the need to force the balance of payments to vanish each period in this way, since gold flows could and did take place under the gold standard. He did, however, require an extra equation to close his model, since he assumed two-bond markets, but included neither of their market-clearing equations in his set of conditions. This shortage presumably induced him to search for some other equation to complete the system, but it also, as stated in Proposition 3 above, leaves these two markets uncleared.

Meade's model under his gold standard assumption is also subject to Proposition 4. Therefore, if his balance-of-payments function is to be consistent with the rest of his model, it must vanish in the second and all subsequent postdisturbance periods, and it then becomes redundant, leaving the model, even on his terms, underdetermined. If on the other hand, the balance of payments does not vanish in this way, the model is contradictory. Moreover, if the balance of payments is set equal to zero in all periods, as Meade's equation seems to state, then no gold could flow between countries even during a first postdisturbance period. This, however, would not allow Meade's money market equations to be satisfied ever, except accidentally.

Later in his book (for example, p. 49) Meade abandons the gold standard assumption and lets the stocks of money be endogenous in his money market-clearing equations. This, however, subjects his model to the criticism presented in Proposition 2. The money equations in this case lock the economies into external balance, which contradicts the results he derives from his balance-of-payments function.

Lastly, Meade includes a net autonomous transfers function, his (1.19). This function must presumably depend on the two countries' excess demands (*ex post*, net purchases) of each other's securities. But, if this is the case, since it is independent of any assets, it must contradict the aggregate budget constraints, as the discussion lying behind Proposition 1 indicates.

While Meade considers a wide range of policies, S. C. Tsiang (1961) essentially utilized Meade's model to investigate the effects of variations in the rate of exchange. He includes a money market equation for each economy ((13) and (14), p. 147) setting each country's demand for that good equal to its supply. He does not include Meade's gold standard equation "as there is hardly any country that mechanically follows this rule of the gold standard game" (p. 147), and he discards Meade's net autonomous transfers function, concentrating only on the balance of trade.

Tsiang sometimes takes the individual money stocks to be exogenous parameters and other times lets them become endogenous by assuming that the monetary authorities vary the quantity of money to either fix the level of income or the interest rate. In either case, we know from Proposition 2 that the balance of payments of each country in this model is forced to equal zero in all periods. However, if the capital markets are segregated (as would seem to be his implicit assumption since each country's equations are independent of the other's rate of interest), this zero value contradicts his balance-of-trade function, which is not required to vanish. As in Meade, Tsiang excludes the bond market equations which again permits these two markets to be out of equilibrium when his equations are satisfied.

J. Marcus Fleming considers the comparative effects for a small country of monetary and fiscal policy under both fixed and flexible exchange rates. His system of equations includes balance-of-trade and net capital import functions and one which sets the demand for money equal to its supply.

Fleming, in his analysis of the effects of an increase in government spending under fixed exchange rates, holds the money stock constant and then derives the conditions for the balance of payments to improve, seemingly unaware that his money market assumption requires the balance of payments

to be zero. In his consideration of the effects of an increase in the supply of money under fixed exchange rates, he continues to require equilibrium in the money market, and then deduces that the equilibrium balance of payments decreases, which is another contradiction.

Jaroslav Vanek (pp. 121–23, 133–37) constructs a standard, small-country Keynesian model with the demand for money set equal to the exogenous supply, which of course should make the value of his balance of payments vanish.

Mundell in a pair of articles (1963, 1964), also considers the effects of monetary and fiscal policies under fixed and flexible exchange rate regimes. In the first part of his 1963 article he verbally considers long-run equilibrium positions in the fixed exchange rate case. In that situation there is asset equilibrium so that the demands and supplies of money are constant over time and equal to each other. There is no contradiction there. However, in the later sections of his paper, Mundell develops a diagrammatic analysis of the problem. There, in the fixed exchange rate case, he uses an FF curve and an LL curve to describe the combination of values of income and the interest rate which bring about external balance and zero excess demand in the money market. He lets increases in the quantity of money and currency flows over the exchanges shift the LL curve, so it obviously holds in the short run. In that case, however, the LL and FF curves should be identical since equations (21)–(26) of Section II above hold whether the economy is in or out of equilibrium. As Mundell's exposition involves a distinction between these two curves, it requires some alteration.

Mundell's first article dealt with a small country faced with perfect capital mobility. His second was meant to generalize the analysis to a situation where the home country is large enough to influence world markets. The model employed is

$$(52)\quad I(r) + \bar{I} - S(Y) + B(Y, Y^*, e) = 0$$

$$(53)\quad I^*(r) - S^*(Y^*) - B(Y, Y^*, e) = 0$$

$$(54)\quad M = L(r, Y)$$

$$(55)\quad M^* = L^*(r, Y^*)$$

$$(56)\quad M = D + R$$

$$(57)\quad M^* = D^* + R^*$$

$$(58)\quad R + R^* = W$$

where $\bar{I}$ is autonomous investment in the home country, D and D^* are the domestic components of the two money stocks, R and R^* are the foreign-reserve components of those stocks, W is the world stock of reserves, and the other variables are as defined in Sections I, II, and III.

Mundell's model is essentially that given by equations (37)–(45) in Section III since his equations (54)–(56) reduce to my equation (39) above.

All of this is unobjectionable, except for two things. The equations of the model are independent of the supply of bonds, which Mundell requires to vary for both his fiscal and monetary policies. This as usual requires the demand function for bonds to be peculiarly specified. Secondly, and more important, the model is a short-run model similar to others used in the literature. As such it, in the first instance, would not seem to provide a complementary analysis to the small-country case which he presents in the first part of his 1963 paper. I say, "in the first instance," because as has been mentioned now several times in this paper, this model reaches its long-run equilibrium position in the second short-run equilibrium period. Therefore, Mundell's analysis of the large-country case is a long-run as well as a short-run analysis. We do note again, however, the lack of realism of such a model.

There is a further criticism of Mundell's analysis in this connection. We have noted that Mundell's conclusions for the fixed exchange rate case are long-run conclusions. Monetary policy, for instance, has no influence in the new long-run equilibrium when there are no longer any money flows between countries.[4] His conclusions for the flexible rate case hold in each short-run period, however, which introduces a

[4] Emphasis on the fact that this is a long-run conclusion is made by Swoboda.

curious asymmetry into the analysis.

In more realistic models than the type Mundell uses, this distinction would be of importance. Monetary policy under fixed exchange rates may be expected to have not inconsiderable effects in the short run before the variations in the stock of money are offset by flows over the exchanges. In the model given by equations (52)-(58) or (37)-(47), however, it makes little difference since the long-run equilibrium position is attained so quickly.

Anne Krueger also analyzes the effect of fiscal and monetary policy under fixed and flexible exchange rates. In equation (10) of her model (p. 197) she sets the demand for money equal to the supply, where the latter variable increases or decreases according to whether government current expenditure less taxes is greater than governmental borrowing. A country's aggregate budget constraint, including all governmental activities other than those of the exchange authority, still implies that the country's balance of payments equals its excess demand for money. Therefore her money market equation also forces the balance of payments to equal zero, which contradicts her balance-of-payments function (2).

Krueger includes a bond market equation (14) and states that it has to hold when the money market equation does. (She presumably meant that it holds when all other markets clear.) She runs into a further difficulty with this formulation, however, for if one writes out her private sector budget constraint, the supply of money enters only the excess demand for money function and the supply of bonds enters only the excess demand function for that good. This involves the model in the sort of contradiction which was discussed in Section I.

Egon Sohmen also considers fiscal and monetary policies under fixed and flexible exchange rate systems. He sets down the following model.

$$(59)\quad Y = A(Y,i,r) + X(r) - M(Y,r) + \eta$$

$$(60)\quad L(Y,i) = \mu + R$$

$$(61)\quad \left[X(r) - \frac{1}{r} M(Y,r)\right] + K(i) = R - R_0$$

where Y is real national income, A is domestic absorption, i is the domestic rate of interest, r is the exchange rate defined as the foreign currency price of domestic currency, X is exports, M is imports, η is government expenditure, R is the domestic currency value of foreign exchange reserves, μ is the autonomous component of the money supply, $L(\cdot)$ is the demand for money, K is capital imports, and R_0 is the initial stock of foreign reserves. With fixed exchange rates, Y, i, and R are the endogenous variables. With flexible rates, r replaces R.

The private sector budget constraint is

$$(62)\quad [Y - A(Y,i,r) + M(Y,r)] - M(Y,r) + K(i) + [\mu + R_0 - L(Y,i)] = 0$$

For the moment we assume government expenditure η is zero. Equations (59) and (62) together imply

$$(63)\quad X(r) - M(Y,r) + K(i) = L(Y,i) - R_0 - \mu$$

However, (63) with equation (60) implies equation (61) above, if $M(Y,r)$ in (59) is multiplied by $1/r$ as it should have been.

The problem here is that (60) and (61) are two ways of expressing the same thing, the excess demand for money. They are not both independent. In addition, neither equation is the equilibrium condition for money in the world economy. Each just determines the amount of money, and therefore reserves, gained or lost by the home economy. Therefore Walras' Law does not permit the omission of the home bond market-clearing equation which should replace either (60) or (61) to give three independent equations.

Reintroducing government expenditure into the model does not alter any of the above, if the reintroduction is done in a consistent way by accounting for the source of funds for that expenditure.

All excess demand functions are independent of the initial stock of bonds which brings about another contradiction if $K(i)$ depends on the excess demand function for bonds as it should do. The final criticism stems from the fact that, although the bud-

get constraint requires (60) and (61) to give the same restriction in equilibrium, they are contradictory. Equation (60) indicates a zero value for the excess demand for money —and therefore for the change in reserves— from the second equilibrium period on following a parameter change, while (61) states that reserves continue to increase period by period.

Harry Johnson uses the *IS-LM* model, essentially as outlined in Section II, but with capital imports added. Therefore, the criticisms stated there hold for his paper.

David J. Ott and Attiat F. Ott use the *IS-LM* model with capital imports and a few other inconsequential variations. Their analysis is also subject to the same criticisms

Ronald W. Jones employs a model with a balance-of-payments function and a money market-clearing equation with the supply of money exogenous. The latter as usual forces the balance of payments to zero which contradicts the former. In addition his model is independent of the stock of bonds, even though these must vary as government expenditure is allowed to alter without changes in the level of taxation or the supply of money.

Jürg Niehans used Sohmen's model. The comments made concerning the latter therefore pertain to Niehans' paper as well.

Richard N. Cooper includes two money market equations (with each country's supply of money equal to the sum of a domestic component and its exchange reserves) which reduce to one which clears the world market for that good. Since, however, he omits both bond market equilibrium conditions, he has one too few equations in his model, a situation he unsuccessfully attempts to circumvent by an argument which results in the increments in the balance of payments being set equal to the changes in the reserves. The omission of the bond market equation also, of course, leaves these two markets uncleared.

Swoboda and Dornbusch set out to investigate the effects of various policies when money flows are incorporated to bring about long-run equilibria. Their first model includes only money and goods and is given by

$$(64)\quad Y = E(Y,\hat{M}) + T(Y,Y',\hat{M},\hat{M}')$$

$$(65)\quad Y' = E'(Y',\hat{M}') - T(Y,Y',\hat{M},\hat{M}')$$

$$(66)\quad M = D + R = L(Y)$$

$$(67)\quad M' = D' + W - R = L'(Y')$$

where unprimed variables refer to the home country and primed to the foreign, Y is income, E is expenditure, T is the trade balance, M is the supply of money, D is the banking system's domestic assets, R is international reserves, L is the demand for money, $\hat{M} = M - L(Y)$ and $\hat{M}' = M' - L'(Y')$. The variables Y, Y', R, $\hat{M}$, $\hat{M}'$, M, and M' are endogenous, and one equation is redundant because of Walras' Law.

The equations are inconsistent. Suppose there is an initial equilibrium solution. Now let D increase. Then the values of Y and Y' which previously satisfied (66) and (67) will no longer do so. However, when these two equations do hold in the new equilibrium, $\hat{M}$ and $\hat{M}'$ are zero in (64) and (65). In that case, however, the previous equilibrium values of Y and Y' will satisfy these two equations, contradicting their new solution values in (66) and (67).

The next model Swoboda and Dornbusch consider is one with segregated capital markets. The equations are

$$(68)\quad Y = (Y,r) + T(Y,Y') + I$$

$$(69)\quad Y' = E'(Y',r') - T(Y,Y') + I'$$

$$(70)\quad M = \overline{D} + R = L(Y,r)$$

$$(71)\quad M' = \overline{D}' + \overline{W} - R = L'(Y',r')$$

$$(72)\quad \frac{dM}{dt} = \frac{dR}{dt} = T(Y,Y') = -\frac{dM'}{dt} = -\frac{dR'}{dt}$$

where in addition to variables defined above, r and r' are the rates of interest in the two countries, and I and I' are shift parameters. Equations (68) through (71) determine Y, Y', r, and r'.

The authors hold M, M', and R constant in each period. Equations (70) and (71) then require each demand for money to equal its

fixed initial supply, and, therefore, if the model is consistent, they force $T(Y, Y')$, the balance of payments, to zero as well. In this case the model reaches its long-run equilibrium position in the first period.

If on the other hand they had let M, M' and R vary, they would not have enough equations to determine the system. This is of course because they have omitted the two bond market equilibrium equations, which in this model are not required to clear. The authors then consider a model with capital mobility which suffers from the same difficulties.

The last paper we consider is by Tsiang (1975), which reconsiders Mundell's problem of appropriate policy assignment. One of the dynamic models Tsaing considers is that of a "free operating" system with no government intervention, (p. 208, equations (17)). The first equation has the time derivative of income proportional to the excess demand for that good. The second has the time derivative of the rate of interest proportional to the excess demand for money. The third has the derivative of the net capital outflow proportional to a function of the four variables of the model, and the last equation has the rate of change of reserves proportional to the balance of payments.

Tsiang takes a linear approximation of the system and then derives the Routh-Hurwitz conditions. The proposition that the balance of payments (defined as receipts less payments) is equal to the excess demand for money is derived from the budget constraint, and it holds in and out of equilibrium. This, however, means that Tsiang's second and fourth equations of the linearized system are proportional if the model is consistent, which makes the determinant of the system equal zero. This, however, contradicts one of the necessary conditions for the system to be stable. Therefore Tsiang's further discussion of the stability conditions is redundant.

VI. Conclusion

The papers discussed in Section V are not on the periphery of the postwar literature. On the contrary, they are established readings in the field. It will be an interesting topic in the history of economic analysis to attempt to explain how international monetary economics was able to develop along these lines for such an extended period of time. However, whatever that explanation is, there is work to be done. Definite results are not easy to obtain in more complex models which do not suffer from contradictions or gross unrealism. That difficulty cannot be sidestepped, however, by constructing the type of analysis which has been typical of the literature.

REFERENCES

R. K. Anderson and A. Takayama, "Devaluation, the Specie Flow Mechanism and the Steady State," *Rev. Econ. Stud.*, June 1977, *44*, 347–61.

R. N. Cooper, "Macroeconomic Policy Adjustment in Independent Economies," *Quart. J. Econ.*, Feb. 1969, *83*, 1–24.

R. Dornbusch, (1973a) "Currency Depreciation, Hoarding, and Relative Prices," *J. Polit. Econ.*, July/Aug. 1973, *81*, 893–15.

——, (1973b) "Devaluation, Money and Nontraded Goods," *Amer. Econ. Rev.*, Dec. 1973, *43*, 871–80.

J. M. Fleming, "Domestic Financial Policies Under Fixed and Floating Exchange Rates," *Int. Monet. Fund Staff Pap.*, Nov. 1962, *9*, 369–79.

J. A. Frenkel and C. H. Rodriguez, "Portfolio Equilibrium and the Balance of Payments. A Monetary Approach," *Amer. Econ. Rev. Proc.*, May 1976, *65*, 163–70.

F. Hahn, "The Balance of Payments in a Monetary Economy," *Rev. Econ. Stud.*, Feb. 1959, *26*, 110–25.

H. G. Johnson, "Theoretical Problems of the International Monetary System," *Pakistan Develop. Rev.*, Spring 1967, *7*, 1–28.

R. W. Jones, "Monetary and Fiscal Policy for an Economy with Fixed Exchange Rates," *J. Polit. Econ.*, July/Aug. 1968, *76*, 921–43.

Murray C. Kemp, "The Rate of Exchange, the Terms of Trade, and the Balance of

Payments in Fully Employed Economies," *Int. Econ. Rev.*, Sept. 1962, *3*, 314-27.

———, *The Pure Theory of International Trade*, Englewood Cliffs 1964.

———, "The Balance of Payments and the Terms of Trade in Relation to Financial Controls," *Rev. Econ. Stud.*, Jan. 1970, *37*, 25-31.

E. A. Kuska, "The Theory of Devaluation, Uniform Commercial Policies, and Transfer Payments," unpublished doctoral dissertation, Univ. London 1970.

———, "The Pure Theory of Devaluation," *Economica*, Aug. 1972, *39*, 309-15.

———, "The Long-Run Behaviour of the Patinkin Model," *Economica*, Aug. 1975, *42*, 292-97.

———, "Devaluation, Equi-proportional Export Subsidies and Import Tariffs, and Transfer Payments," *Economica*, May 1976, *43*, 182-84.

———, "The Post-Devaluation Time Profile of Reserves and Prices under Neoclassical Assumptions," *Economica*, Aug. 1977, *44*, 289-92.

P. J. K. Kouri and M. G. Porter, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May/June 1975, *82*, 443-46.

A. O. Krueger, "The Impact of Alternative Government Policies Under Varying Exchange Rates," *Quart. J. Econ.*, May 1965, *79*, 195-208.

John F. Kyle, *The Balance of Payments in a Monetary Economy*, Princeton 1976.

James E. Meade, *The Theory of International Economic Policy, Vol. I: The Balance of Payments, Mathematical Supplement*, London 1951.

R. A. Mundell, "Capital Mobility and Stabilization Policy Under Fixed and Flexible Exchange Rates," *Can. J. Econ.*, Nov. 1963, *29*, 475-85.

———, "A Reply: Capital Mobility and Size," *Can. J. Econ.*, Aug. 1964, *30*, 421-31.

M. Mussa, "A Monetary Approach to Balance-of-Payments Analysis," *J. Money, Credit, Banking*, Aug. 1974, *6*, 333-52.

T. Negishi, "Approaches to the Analysis of Devaluation," *Int. Econ. Rev.*, June 1968, *9*, 218-27.

J. Niehans, "Monetary and Fiscal Policies in Open Economies Under Fixed Exchange Rates: An Optimizing Approach," *J. Polit. Econ.* July/Aug. 1968, *76*, 893-920.

O. J. Ott and A. F. Ott, "Monetary and Fiscal Policy: Goals and the Choice of Instruments," *Quart. J. Econ.*, May 1968, *82*, 313-25.

Don Patinkin, *Money, Interest and Prices*, 2d ed., New York 1965.

E. Sohmen, "Fiscal and Monetary Policies Under Alternative Exchange Systems," *Quart. J. Econ.*, Aug. 1967, *81*, 515-23.

A. K. Swoboda, "Monetary Policy Under Fixed Exchange Rates: Effectiveness, the Speed of Adjustment and Proper Use," *Economica*, May 1973, *40*, 136-54.

——— **and R. Dornbusch,** "Adjustment Policy and Monetary Equilibrium in a Two-Country Model," in Michael B. Connolly and Alexander K. Swoboda, eds., *International Trade and Money*, London 1973.

S. C. Tsiang, "The Role of Money in Trade-Balance Stability, Synthesis of the Elasticity and Absorption Approaches," *Amer. Econ. Rev.*, Dec. 1961, *51*, 912-36.

———, "The Dynamics of International Capital Flows and Internal and External Balance," *Quart. J. Econ.*, May 1975, *89*, 195-214.

Jaroslav Vanek, *International Trade: Theory and Economic Policy*, Homewood 1962.

Dynamic Stability and the Theory of Factor-Market Distortions

By J. PETER NEARY*

Although international monetary economists have devoted much attention to the process of adjustment from one equilibrium to another, the pure theory of international trade has traditionally confined its attention to comparisons between long-run equilibria.[1] The same point applies to those branches of theory, such as the neoclassical theory of tax incidence, which make use of models identical in structure to the Heckscher-Ohlin-Samuelson model of international trade. The aim of this paper is to suggest that this neglect has overlooked many interesting problems, and to argue that the study of adjustment mechanisms in two-sector neoclassical models is both of interest in itself, and of value in providing information on the comparative static properties of these models. In particular, it is shown that the use of explicit adjustment mechanisms permits some considerable simplifications of the theory of factor-market distortions

Many of the recent writings on factor-market distortions by international trade theorists have been concerned with the elucidation of a number of paradoxes (paradoxes at least in the light of accepted trade theory) which can arise in the presence of such distortions. (See especially Jagdish N. Bhagwati and T. N. Srinivasan; Ronald W. Jones, 1971a; Stephen P. Magee, 1971, 1976.) One writer has gone as far as to say that the introduction of factor-market distortions into international trade models opens a "Pandora's Box of paradoxes" (see Raveendra N. Batra, p. 279). A necessary and sufficient condition for the occurrence of many of these paradoxes is that the economy exhibit a particular condition (which cannot arise if factor markets are undistorted): namely, the level of initial distortions must be such that the ranking of the two sectors in terms of *physical* factor intensities is the opposite of their ranking in terms of *value* factor intensities; in other words, in the sector which uses a higher physical ratio of capital to labor, the share of payments to capital in the total value of output is lower than in the other sector. The implications of this condition may be seen by noting the principal paradoxes to which it gives rise:

1) *Perverse Price Output Response*: An increase in the relative price of one good leads to a fall in its output, assuming the levels of factor-market distortions are unchanged.

2) *Perverse Distortion-Output Response*: An increase in the rate of subsidy to one sector (whether an output or an input subsidy) leads to a fall in the output of that sector, assuming relative output prices are unchanged.

3) *Lack of Correspondence between Rybczynski and Stolper-Samuelson Theorems*: Each of these theorems continues to hold in isolation, but the former must be expressed in terms of physical factor intensities, and the latter in terms of value factor intensities. Hence, if labor force growth at constant commodity prices increases the output of good X, an increase in the relative price of X assuming a constant labor force will *reduce* rather than increase the real wage. More surprisingly still, a country may be capital abundant relative to the rest of the world, and exporting its

*Heyworth research fellow, Nuffield College, Oxford. I am indebted to Nick Stern for detailed comments on earlier drafts, and to George Borts, Dermot McAleese, and an anonymous referee for helpful suggestions.

[1]Since this paper was written, my attention has been drawn to two interesting papers, by Wolfgang Mayer (1974a) and by Murray C. Kemp, Yoshio Kimura, and Koji Okuguchi, which study adjustment mechanisms similar to those considered here. However both of these papers are exclusively concerned with a small open economy, and do not discuss the question of factor-market distortions.

physically capital-intensive commodity, with all the other conditions for the Heckscher-Ohlin theorem holding, and yet a protection-induced increase in the domestic price of the import-competing good will *reduce* the real return of the scarce factor (labor).

These examples are sufficient to demonstrate that international trade theory becomes apparently much more complicated when allowance is made for factor-market distortions. However, a major object of this paper is to show that *all these paradoxes are theoretical curiosa which will "almost never" be observed in real world economies*. More formally, Sections I and II below demonstrate that in a small open economy, equilibria where the value and physical factor-intensity rankings of the two sectors differ are necessarily *unstable* under a variety of plausible adjustment mechanisms.

Section I concentrates on the "short-run capital specificity" adjustment process, which is one exception to the general neglect of adjustment processes by international trade theorists, having been studied by Mayer (1974b) and Michael Mussa, drawing on earlier work by Jones (1971b). A more general class of adjustment mechanisms, permitting both labor and capital to be sector specific in the short run, is examined in Section II. Section III extends the analysis to the case where relative commodity prices are endogenous and derives a general stability condition for this case also. The implications of these findings are examined in Section IV: it is shown that they permit a considerable simplification of both international trade theory and the general equilibrium theory of tax incidence, and also that they may be interpreted in terms of the relationship between Walrasian and Marshallian stability conditions. Finally the principal conclusions of the paper are summarized in Section V.

I. Short-Run Capital Specificity in a Small Open Economy

In this section, I examine the process of adjustment to an exogenous change in the level of factor-market distortion in a small open economy which obeys all the assumptions of the Heckscher-Ohlin-Samuelson model of international trade (including perfect competition, constant returns to scale and fixed aggregate factor supplies) with two exceptions: firstly, long-run equilibrium is characterized by constant proportional differentials between the value marginal products of capital and labor in the two sectors:[2]

(1a) $$w_Y = \alpha w_X$$

(1b) $$r_Y = \beta r_X$$

and secondly, while the labor force is instantaneously reallocated to ensure that (1a) is satisfied at all times, capital goods are sector specific in the short run, and move between sectors in the medium run in response to deviations from (1b).[3]

To investigate the short-run effect on the intersectoral differential in capital rentals of a once and for all increase in the labor-market distortion parameter α (attributable, for example, to an increase in the rate of labor subsidy to sector X), I invoke the well-known conditions, implied by profit maximization and constant returns to scale, that the proportional change in the price of each good must equal a weighted average

[2]The term long run is to be understood in the Marshallian sense, as a period long enough for each sector to have adopted a scale of production which is (privately) optimal in the light of its external environment. A corollary of this is that factors have been allocated such that they earn the same net return in each sector. Since throughout this paper I assume that total factor supplies are fixed, it is clear that this long run may be relatively short in comparison with the time periods usually considered in the theory of economic growth. The sense in which the appellation "Marshallian" is appropriate should become clear in Section IV below.

[3]Most writers in the enormous literature on factor-market distortions deal exclusively with the long-run, two mobile factors, case. (See especially Arnold C. Harberger and Jones, 1971a.) Exceptions include Charles E. McLure, Jr. who uses a model where labor is sector specific but capital instantaneously mobile. Magee (1976, pp. 85–86) who briefly discusses the adjustment process, citing unpublished work by J. Marquez-Ruarte; and the author (1978), who examines the consequences of a particular factor-market distortion in both mobile and immobile capital models, but without examining the adjustment process.

of the changes in factor prices in each sector, the weights being the share of each factor in the value of output of that sector (a circumflex indicates a proportional rate of change: $\hat{w} = d\log w$):

$$\hat{p}_X = \theta_{LX}\hat{w}_X + \theta_{KX}\hat{r}_X \tag{2}$$

$$\hat{p}_Y = \theta_{LY}\hat{w}_Y + \theta_{KY}\hat{r}_Y \tag{3}$$

With both commodity prices held constant because of the "small country" assumption, equations (1a), (2), and (3) may be manipulated to express the change in the intersectoral rental differential as a function of changes in the labor-market distortion parameter and the wage in the X sector:

$$\theta_{KX}\theta_{KY}(\hat{r}_X - \hat{r}_Y) = -|\theta|\,\hat{w}_X + \theta_{LY}\theta_{KX}\hat{\alpha} \tag{4}$$

where $|\theta|$ is the determinant of the matrix of sectoral factor shares, and is positive if and only if the X sector is relatively labor intensive in the *value* sense. To eliminate the change in w_X from (4), use the full-employment condition, which is expressed by equating the sum of the labor demand schedules of each sector to the fixed labor endowment:

$$L_X\left(\frac{w_X}{p_X}, K_X\right) + L_Y\left(\frac{w_Y}{p_Y}, K_Y\right) = \bar{L} \tag{5}$$

Differentiating (5), holding commodity prices and capital allocations constant, and using equation (1a), one may express the short-run proportional change in the wage rate in sector X as a function of the change in the labor market distortion parameter.[4]

$$\hat{w}_X = -\frac{1}{\Delta}\lambda_{LY}\frac{\sigma_Y}{\theta_{KY}}\hat{\alpha} \tag{6}$$

where

$$\Delta = \lambda_{LX}\frac{\sigma_X}{\theta_{KX}} + \lambda_{LY}\frac{\sigma_Y}{\theta_{KY}} > 0 \tag{7}$$

Finally, substituting from (6) into (4):

$$\theta_{KX}\theta_{KY}(\hat{r}_X - \hat{r}_Y) = \frac{1}{\Delta}(\sigma_X\lambda_{LX}\theta_{LY} + \sigma_Y\lambda_{LY}\theta_{LX})\hat{\alpha} \tag{8}$$

Assuming that the capital market is in long-run equilibrium before the change, this shows that an increase in the wage differential in favor of sector Y will increase the relative rental on capital in the other sector in the short run. Hence, in the medium run, capital will begin to reallocate from sector Y to sector X, *irrespective of the relative factor intensities* (in either the physical or value sense) of the two sectors.

But this capital reallocation will itself affect the intersectoral rental differential. To see this, again differentiate equation (5), this time holding commodity prices and factor-market distortions constant:

$$\hat{w}_X = \frac{1}{\Delta}(\lambda_{LX}\hat{K}_X + \lambda_{LY}\hat{K}_Y) \tag{9}$$

This gives the effect on the X sector wage rate of any arbitrary changes in the capital stocks in each sector. However, when a reallocation of the given stock of capital is considered, the changes in each sector are not arbitrary, but must continue to satisfy the capital endowment constraint:

$$K_X + K_Y = \bar{K}$$

or in differential form:

$$\lambda_{KX}\hat{K}_X + \lambda_{KY}\hat{K}_Y = 0 \tag{10}$$

Substituting from (10) into (9) gives

$$\hat{w}_X = \frac{|\lambda|}{\Delta}\frac{\hat{K}_X}{\lambda_{KY}} \tag{11}$$

where $|\lambda|$ is the determinant of the matrix of factor-to-sector allocations, and is positive if and only if sector X is relatively labor intensive in the *physical* sense. Finally, substituting from (11) into (4), holding constant the level of the labor-market distortion parameter α, one derives

$$\hat{r}_X - \hat{r}_Y = -\frac{1}{\Delta}\frac{|\lambda|\,|\theta|}{\theta_{KX}\theta_{KY}}\frac{\hat{K}_X}{\lambda_{KY}} \tag{12}$$

This shows that a reallocation of capital into

[4]The notation used is that of Jones (1965). λ_{ij} is the proportion of the fixed stock of factor i used in sector j; θ_{ij} is the share of gross payments to factor i in the value of output of sector j, and σ_j is the elasticity of substitution between capital and labor in sector j.

sector X will reduce the proportional gap between the rentals in the two sectors, if and only if the product of the two determinants, $|\lambda|$ and $|\theta|$, is positive; in other words, if and only if the rankings of the two sectors by physical and value factor intensities are the same.[5] Hence, if these rankings differ, the initial reallocation of capital in response to an increase in the wage differential will widen the intersectoral rental differential, leading to an increased flow of capital from Y to X, moving progressively further away from the new long-run equilibrium predicted by the comparative static analysis One may conclude, therefore, that under the short-run capital specificity adjustment process, an equilibrium where the physical and value factor intensity rankings differ in a small open economy must be locally unstable.

II. Stability in a Small Open Economy when Both Capital and Labor Adjust Sluggishly

In the last section I examined the stability of equilibrium following a particular exogenous change (a change in the level of distortion in the labor market), and assuming a particular adjustment mechanism (the short-run capital-specificity process). In this section I generalize this to the case where both capital and labor adjust sluggishly, and show that the same stability criterion applies. I continue to adopt the small open economy assumption of fixed relative commodity prices.

The dynamic assumptions of this section are that both capital and labor follow adjustment processes of the same form (though not necessarily of the same speed). Both factors are now sector specific in the short run, but complete flexibility of the wage and rental rate within each sector ensures that both factors are fully employed throughout the adjustment period.[6] As before, a long-run equilibrium is defined as a state where each factor earns the same return in both sectors, taking into account the distortion parameters α and β as in equations (1a) and (1b). Starting from such an equilibrium, any exogenous shock will bring about a short-run intersectoral divergence of the distortion-inclusive returns of each factor. These divergences in turn will lead in the medium run to a gradual reallocation (or "migration") of each factor out of the sector where it earns a lower return into the sector where it earns a higher return. The rate at which this reallocation takes place will depend in general on a variety of considerations, including migration costs, rates of time preference and expectations of future price changes on the part of factor owners. However, since our primary interest lies not in these influences themselves, but rather in their implications for stability, we may subsume them into general migration functions of the following form (where D is the time derivative operator):

$$(13)\quad DK_X = \phi\left\{\beta\frac{r_X}{r_Y} - 1\right\} \qquad \phi' > 0, \phi(0) = 0$$

$$(14)\quad DL_X = \psi\left\{\alpha\frac{w_X}{w_Y} - 1\right\} \qquad \psi' > 0, \psi(0) = 0$$

It turns out that these dynamic assumptions may be conveniently analyzed by lo-

[5]This point has been noted by Jones (1975, p 13), who nevertheless claimed that the adjustment process would eventually converge because, as capital reallocates, the *absolute* gap between the two rentals is narrowed However, he did not point out that the adjustment is taking place in the opposite direction to that predicted by the comparative static analysis.

[6]One difficulty with this assumption arises if either factor is "inessential" in the production function of either sector (in the sense that the isoquants of the function intersect the axis corresponding to that factor) In such cases the assumption of complete factor-price flexibility implies that the return of the relevant factor in the sector concerned may be zero during part of the adjustment period This situation is logically consistent, provided it is assumed that the inessential factor is maintained above subsistence level by lump sum transfers where necessary. However, it is not a very appealing situation, and where it is likely to occur a different adjustment mechanism, similar to those mentioned in fn 12 below, may be preferred to that of the present section.

cating directly in the Edgeworth-Bowley box the stationary loci of (13) and (14)[7] (i.e., equations (1a) and (1b)) as well as the distorted efficiency locus (*DEL*) or contract curve of the Edgeworth-Bowley box, whose equation is

$$\frac{\alpha}{\beta}\frac{w_X}{r_X} = \frac{w_Y}{r_Y} \tag{15}$$

The details of this construction are outlined in the Appendix, Section A. Figure 1 presents the case where sector X is capital intensive in the physical sense ($|\lambda| < 0$), and where the rankings of the two sectors by physical and value factor intensities correspond ($|\lambda|\ |\theta| > 0$). For all points other than the long-run equilibrium point E, the directions of movement are shown by the arrows. These may be verified by reference to equations (A1) and (A3) in the Appendix, and are in any case intuitively plausible. For example, an upward movement away from any point on the capital-market equilibrium locus (*KMEL*) implies an increase in the capital-labor ratio in X and a decrease in that in Y; at constant commodity prices this implies a fall in the rental in X and an increase in that in Y; hence capital is encouraged to move out of the X sector into Y. The direction of labor migration at all points off the labor-market equilibrium locus (*LMEL*) may be established in a similar manner. The conclusion to be drawn from the figure is that the equilibrium at E is globally as well as locally stable.

A very different conclusion is drawn from Figure 2, however, where the physical and value factor-intensity rankings of the two sectors differ. At the equilibrium point F, the *KMEL* cuts the *DEL* from below, while

FIGURE 1

the *LMEL* cuts it from above. As a result, the point F is a saddle point: there is one knife-edge path, indicated by the dotted line JJ', along which factor allocations will converge towards F. But this path is itself unstable: the slightest divergence from it will lead to a cumulative movement away from F. Hence, while the condition $|\lambda|\ |\theta| < 0$ does not precisely imply global instability, equilibria where this condition is met are likely in practice to be highly unstable.

Two special cases of the general adjustment mechanism may be mentioned in passing. The first, where the labor force is assumed to be reallocated instantaneously, is simply the short-run capital specificity adjustment process of Section 1. Only points

FIGURE 2

[7]The only previous writer to discuss these loci appears to be Magee, who considered only one of the two loci, calling it first a "distortion equilibrium locus" (1971, p. 630), and later, following Marquez-Ruarte, an "iso-price locus" (Magee, 1976, pp. 28, 86). The latter term is inappropriate in the present context, however, since with both factor markets in disequilibrium, any point in the production box is a feasible short-run equilibrium for a given relative goods price

along the *LMEL* in Figures 1 and 2 are now admissible, and it is clear that point *E* in Figure 1 is still globally stable while point *F* in Figure 2 is now globally unstable, thus reinforcing the conclusions of Section I. A second special case is where capital goods are sector specific in the long run (or, more plausibly, "capital" in each sector is a different fixed factor, for example, plant and machinery in the manufacturing sector *X*, and land in agriculture *Y*, as in Jones, 1971b). In this case, it may be seen that no instability problem arises, even in Figure 2. We may conclude therefore that the comparative static paradoxes and the associated problem of instability arise only when *both* capital and labor are assumed to be intersectorally mobile in the long run.

Finally, having shown that, when the rankings of sectors by physical and value factor intensities differ, the economy will not move towards the new equilibrium predicted by the comparative static analysis, one must establish where it will tend towards. In brief, the answer is either towards a *specialized* equilibrium, or towards a new nonspecialized equilibrium where the initial ranking of the sectors by *value* factor intensity has been reversed, so that the physical and value rankings now coincide. To illustrate this, refer again to Figure 2. As the diagram is drawn, all paths other than those starting along the line *JJ'* converge towards either O_X or O_Y, implying that the economy specializes in good *Y* or good *X*, respectively. Nevertheless, it is possible that the process of factor reallocation may reverse the value factor-intensity ranking of the sectors, in which case a new stable long-run equilibrium may be attained at some point other than *F* on the *DEL*. At such a point, both the *KMEL* and the *LMEL* would again intersect the *DEL*, and a local stability analysis similar to that of Figure 1 would be appropriate. However, if this reversal does not occur (for example, it cannot occur if both production functions are Cobb-Douglas), then the factor reallocation will continue until the economy specializes in the production of one or other good.

III. Preexisting Distortions and Stability of Equilibrium when Commodity Prices are Variable

So far I have been exclusively concerned with a small open economy, where prices are parametrically given. However, the adjustment processes just considered may be extended to examine the stability of equilibrium when commodity prices are variable. To investigate this, follow Jones (1965, 1971a) and assume that the aggregate demand function, in differential form, is given by

$$(16) \qquad \hat{X} - \hat{Y} = -\sigma_D(\hat{p}_X - \hat{p}_Y)$$

This may be interpreted as referring to a closed economy where aggregate preferences are homothetic, in which case the parameter σ_D is the elasticity of substitution in demand. Equation (16) could alternatively be interpreted as a composite demand function, representing the combined effects of home and foreign demand in an open economy which possesses some monopoly power in international trade. In both cases, assume that σ_D is positive.

In addition to specifying the demand function (16), assume that relative commodity prices move instantaneously to clear the commodity market at all times. When this assumption is made, it is shown in the Appendix, Section B that a necessary and sufficient condition for local stability is

$$(17) \quad \frac{1}{\sigma_E}(\sigma_D|\lambda|\,|\theta| + \sigma_X Q_X + \sigma_Y Q_Y) > 0$$

where $Q_X = \lambda_{LX}\theta_{KX} + \lambda_{KX}\theta_{LX}$

and $Q_Y = \lambda_{LY}\theta_{KY} + \lambda_{KY}\theta_{LY}$

Hence, the stability condition (assuming that σ_d is positive) is that the expression in parentheses in (17) be positive. This expression, denoted by σ, has been termed by Jones (1965) the economy's "aggregate elasticity of substitution." Its importance for the comparative static results of the two-sector model is well known, but its crucial role in affecting stability does not appear to have been noted by previous writers. Note that the stability criterion for

the fixed-prices model of Sections I and II is a special case of (17), corresponding to the limiting value as σ_D tends to infinity.

While the local stability criterion $\sigma > 0$ is just as convenient as that for the small open-economy model $|\lambda|\ |\theta| > 0$, the analysis of global stability is more complicated when prices are variable, since there is no guarantee that either factor-market equilibrium locus will slope upwards. There are now three configurations which the phase diagram may exhibit in the neighborhood of a long-run equilibrium point: (i) both loci may slope upwards; in this case the analysis of Figures 1 and 2 is directly applicable: the equilibrium will be locally stable if and only if the *KMEL* cuts the *DEL* from above and the *LMEL* cuts it from below; (ii) one locus may be upward sloping and the other downward sloping; in this case condition (17) is necessarily fulfilled and the equilibrium is locally stable; or (iii) both loci may be downward sloping; here (17) requires that the slope of the *KMEL* be greater (i.e., less negative) than that of the *LMEL* for local stability. Finally, as in Section II, it is quite possible for multiple equilibria to exist, with the two equilibrium loci crossing and recrossing the *DEL*, implying a succession of alternately stable and unstable equilibria.

IV. Implications of the Stability Criteria

Evidently, the importance of obtaining stability conditions derives from the fact that they may be combined with Samuelson's Correspondence Principle (see ch. 9), to assert that, since all empirically interesting equilibria must be stable, long-run equilibria where the stability conditions do not hold will "almost never" be observed.

The implications of the stability condition $|\lambda|\ |\theta| > 0$ for the theory of factor-market distortions in an open economy are immediately obvious: all of the comparative static paradoxes mentioned in the introduction become inconsistent with stable unspecialized equilibrium. However, it is important not to claim too much for this stability condition. For example, as Bhagwati and Srinivasan and others have shown, there is no necessary relationship between the reversal of physical and value factor intensities and the curvature of the distortion-constrained transformation curve. Hence, the stability condition $|\lambda|\ |\theta| > 0$ does not imply that the transformation curve must be concave to the origin in the neighborhood of an equilibrium point. Nor does it rule out the possibility that international factor price equalization may be prevented by the existence of factor-market distortions. Furthermore, there are a number of "paradoxes" which are attributable to factor-market distortions but which can occur even when the physical and value factor-intensity rankings are the same.[8] Clearly, nothing in the present paper rules out the possibility of any of these paradoxes.

Turning to the stability criterion for a closed economy $\sigma > 0$, it is of considerable importance, for the aggregate elasticity of substitution is the key parameter in most of the comparative static derivations for the two-sector model. For example, as shown by Jones (1965, equation (15)), an increase in the aggregate capital-labor ratio will raise the wage rental ratio if and only if σ is positive: our stability condition therefore rules out "capital-intensity perversities" in the two-sector model with homogeneous capital. It can also be shown (see the author, 1976) that provided σ is positive the conclusions of Harberger concerning factor-tax incidence in a closed economy and those of Harry G. Johnson and Peter Mieszkowski concerning the income distribution effect of unionization continue to hold in a wide

[8] For example if a differential is paid by the import-competing sector, free trade may be inferior to no trade (see Batra, p. 264); if it is paid by the export sector, an increase in the terms of trade may reduce welfare (Batra, pp. 268–70); and growth may be immiserizing even when the terms of trade are constant and there are no impediments to trade (Batra, pp. 272–73). As Jagdish N. Bhagwati has pointed out, most if not all of these paradoxes may be viewed as special cases of the general phenomenon of immiserizing growth.

variety of circumstances.[9] Finally, this condition may be related to the two-sector neoclassical growth model of Hirofumi Uzawa: a necessary and sufficient condition for uniqueness of momentary equilibrium in this model is that σ be positive, as Avinash K. Dixit and others, following Frank H. Hahn, have shown. The approach adopted here therefore draws attention to the unity of the two-sector framework, whether applied to the theories of public finance, international trade, or economic growth.[10]

A rather different application of this stability condition is the light it throws on the relationship between the so-called Marshallian and Walrasian stability criteria. For example, in examining closed-economy equilibria where the aggregate supply schedule slopes downwards, Jones (1971a) assumed that its intersection with the aggregate demand schedule should exhibit Walrasian stability. Such an intersection is shown at C_o in Figure 3: the downward-sloping long-run supply schedule S_o cuts the demand schedule D from above, implying that excess demand is positive for prices just below and negative for prices just above the equilibrium price p_o.

Is the stability criterion $\sigma > 0$ satisfied at C_o in Figure 3? It is easily seen that it is not. For at that point, the algebraic value of the elasticity of the aggregate demand schedule is less than that of the long-run aggregate supply schedule, or in symbols: $-\sigma_D < \sigma_{LRS}$. Invoking the expression for the long-run supply elasticity derived by Jones (1965), this may be rewritten as

FIGURE 3

$$\sigma_D + \frac{\sigma_X Q_X + \sigma_Y Q_Y}{|\lambda|\,|\theta|} > 0 \tag{18}$$

But, since the aggregate supply curve is downward sloping, the term $|\lambda|\,|\theta|$ is negative. Hence, multiplying (18) by $|\lambda|\,|\theta|$ reverses the sign of the inequality:

$$\sigma_D|\lambda|\,|\theta| + \sigma_X Q_X + \sigma_Y Q_Y < 0$$

This shows that the equilibrium at C_o is indeed unstable in the long-run sense. For example, the granting of a subsidy to sector X shifts the long-run supply curve from S_o to S_1, implying that the new long-run equilibrium should be at E (note that the variability of commodity prices does not affect the comparative static prediction of a perverse distortion-output response in the long run). But the adjustment path which the economy actually follows involves a sequence of short-run equilibria, moving down the demand curve in the direction of the double-headed arrow, and converging towards a new stable long-run equilibrium at G. At this point the long-run supply curve cuts the demand curve from below, implying that the stability condition $\sigma > 0$ is indeed fulfilled there. Hence, contrary to Jones, the correct criterion for the stability of long-run equilibrium in this model is that the aggregate supply and demand schedules

[9]In discussing their principal equation, Johnson and Mieszkowski (p. 550) invoke stability considerations in order to establish the sign of the denominator (which may be shown to be a simple multiple of σ). However they do not specify a dynamic adjustment process, and other writers on tax incidence in the presence of nonzero initial taxes have ignored this issue (see, for example, Adolf I. Vandendorpe and Ann F. Friedlaender, p. 219).

[10]The parallel with growth theory is not exact however, since I am comparing a static model exhibiting short-run factor specificity with a dynamic model where capital is instantaneously transferable between sectors. For studies of two-sector growth models with capital specificity, see Ken-Ichi Inada and Antonio Bosch, Andreu Mas-Colell, and Assaf Razin.

exhibit Marshallian, not Walrasian, stability.[11]

Does this mean that there is no role for the Walrasian stability criterion? The answer to this is no, for the model must also exhibit Walrasian stability, but this refers to the intersection of the demand curve with the *short-run* supply curve. A different short-run supply curve is implied by each of the adjustment mechanisms considered. For example, if, as in Sections II and III, both capital and labor are sector specific in the short run but remain fully employed, the short-run supply curve is vertical. Alternatively, under the adjustment mechanism assumed in Section I, the short-run supply curve is upward sloping, as shown in Figure 3. In the latter case, the effect of increasing the rate of subsidy in the X sector is to shift the short-run supply curve to the right, leading to a new equilibrium at C_1. But, irrespective of the adjustment mechanism assumed, as factor markets move towards long-run equilibrium, the short-run supply curve shifts progressively to the right, tracing out a sequence of short-run equilibria where each short-run supply curve intersects the demand curve. At each of these intersections Walrasian stability must prevail. It should be noted that this analysis is perfectly consistent with the usual explanation for the difference between Marshallian and Walrasian stability (see Mark Blaug, pp. 411–14): that the former assumes it is quantities which adjust out of equilibrium whereas the latter assigns that role to prices. The present paper confirms this analysis, and supplements it by showing that under the adjustment mechanisms considered above, the two stability criteria are complementary; since in full long-run equilibrium both must hold, though the Marshallian criterion must be interpreted with respect to the long-run supply curve, whereas the Walrasian criterion is applicable to the short-run supply curve.

[11]The implicit assumption made by Jones is that the commodity market is cleared by a tatonnement price adjustment mechanism, but that factor markets adjust completely to each suggested price announced by the hypothetical auctioneer. But in this bizarre situation, the model would be unstable whenever the aggregate supply schedule were downward sloping, since factor markets would be adjusting in the face of fixed commodity prices (even though no trading was actually taking place) and so the stability analysis of Section II would be appropriate.

V. Summary and Conclusion

This paper has examined the implications for the theory of factor-market distortions of a class of dynamic adjustment mechanisms, of which the short-run capital specificity assumption is an interesting special case. It was shown that the existing literature requires modification in a number of respects:

1) A small open economy (i.e., an economy facing fixed commodity prices) can *never* be in a stable unspecialized equilibrium where one sector is relatively labor intensive in physical terms, but relatively capital intensive in value terms. This conclusion rules out a number of paradoxical outcomes (such as a "perverse" price-output response) which have received a great deal of attention in the literature on factor-market distortions.

2) A closed economy, or an open economy which has some influence over its terms of trade, *can* be in a stable unspecialized equilibrium where the rankings of the two sectors by physical and value factor intensities do not correspond. The condition for such an equilibrium to be stable is that Jones' "aggregate elasticity of substitution" be positive. This stability condition is important since it permits us to sign unambiguously a great many comparative static results of the two-sector model; and in this case too, many paradoxical outcomes may be ruled out.

3) Finally, contrary to Jones (1971a), when the rankings of the two sectors by physical and value factor intensities differ, the condition for the economy to be in stable equilibrium implies that the intersection of the demand schedule and the long-run supply schedule must exhibit Walrasian instability.

One final qualification to the conclusions of this paper must be noted. It involves

what might be called the Ambiguity of the Correspondence Principle: while the assumption of stability is a powerful tool in deducing comparative static results, it may yield different results depending on which particular disequilibrium adjustment mechanism is assumed. Thus all of the conclusions given above are conditional on the particular class of adjustment mechanisms which has been examined. However, while it is not difficult to devise alternative mechanisms, there is no reason to believe that they would imply different stability conditions.[12]

Clearly, it would be desirable to develop still more complicated adjustment mechanisms, insofar as the objective is to model the actual process of adjustment towards long-run equilibrium in real world economies.[13] But for the present it seems reasonable to accept the conclusions derived from the adjustment mechanisms studied in this paper, with their attractive implication that the long-run predictions of the theory of factor-market distortions are much more consistent with simple economic intuition than has been thought.

Appendix

A

This section derives the properties which must be satisfied in Figures 1 and 2 by the three loci, the labor-market equilibrium locus (*LMEL*) equation (1a), the capital-market equilibrium locus (*KMEL*) equation (1b), and the distorted efficiency locus (*DEL*) equation (15). The procedure adopted is to totally differentiate each locus, and then to convert each to a differential function of changes in relative commodity prices and factor allocations to sector X only. The latter is accomplished by invoking: (a) the price-equal-to-unit-cost equations, (2) and (3); (b) the relationship between the capital-labor ratio and the marginal product of capital in sector X, in differential form,

$$\text{(A1)} \quad \hat{K}_X - \hat{L}_X = -\frac{\sigma_X}{\theta_{LX}}(\hat{r}_X - \hat{p}_X)$$

as well as the corresponding equation for sector Y; and (c) the full-employment constraints: equation (10) for capital, and for labor:

$$\text{(A2)} \quad \lambda_{LX}\hat{L}_X + \lambda_{LY}\hat{L}_Y = 0$$

This procedure yields the following expressions for the three loci.

$$\text{(A3)} \quad KMEL.\ \hat{p}_X - \hat{p}_Y - A_1\hat{K}_X + B_1\hat{L}_X = 0$$

$$\text{(A4)} \quad A_1 = \frac{\theta_{LX}}{\sigma_X} + \frac{\theta_{LY}}{\sigma_Y}\frac{\lambda_{KX}}{\lambda_{KY}}$$

$$\text{and} \quad B_1 = \frac{\theta_{LX}}{\sigma_X} + \frac{\theta_{LY}}{\sigma_Y}\frac{\lambda_{LX}}{\lambda_{LY}}$$

$$\text{(A5)} \quad LMEL:\ \hat{p}_X - \hat{p}_Y + A_2\hat{K}_X - B_2\hat{L}_X = 0$$

$$\text{(A6)} \quad A_2 = \frac{\theta_{KX}}{\sigma_X} + \frac{\theta_{KY}}{\sigma_Y}\frac{\lambda_{KX}}{\lambda_{KY}}$$

$$\text{and} \quad B_2 = \frac{\theta_{KX}}{\sigma_X} + \frac{\theta_{KY}}{\sigma_Y}\frac{\lambda_{LX}}{\lambda_{LY}}$$

$$\text{(A7)} \quad DEL:\ A_3\hat{K}_X - B_3\hat{L}_X = 0$$

$$\text{(A8)} \quad A_3 = \frac{1}{\sigma_X} + \frac{1}{\sigma_Y}\frac{\lambda_{KX}}{\lambda_{KY}}$$

$$\text{and} \quad B_3 = \frac{1}{\sigma_X} + \frac{1}{\sigma_Y}\frac{\lambda_{LX}}{\lambda_{LY}}$$

Assuming commodity prices and the distortion parameters to be fixed, we now have three loci to locate in the Edgeworth-Bowley box, with the elasticity of each given by $B_i/A_i (i = 1, 3)$. We may note first that at any point of long-run equilibrium all three must intersect; moreover, at such a point the difference between their slopes will have the same sign as the difference between their elasticities. Hence, by comparing the elasticities of *KMEL*, *LMEL*, and *DEL*, the follow-

[12]It may be shown that the stability conditions of this paper continue to apply under alternative adjustment mechanisms, which allow for short-run labor market segmentation and sluggish adjustment of wage levels in response to excess labor supply

[13]Among the least satisfactory assumptions of the present model that may be mentioned are the absence of any explicit costs to factor reallocation, the assumption that elasticities of substitution in production are no greater in the long run than in the short run, and the lack of consideration given to monetary factors. Moreover, it may be questioned whether the process of capital reallocation can be distinguished even conceptually from that of capital accumulation

ing relationships between the slopes of the three loci, at any point of long-run equilibrium, may be established:

$$\text{(A9)}\quad |\lambda|\,|\theta| > 0 : 0 < \left.\frac{dK_X}{dL_X}\right|_{KMEL} < \left.\frac{dK_X}{dL_X}\right|_{DEL} < \left.\frac{dK_X}{dL_X}\right|_{LMEL}$$

$$\text{(A10)}\quad |\lambda|\,|\theta| < 0 : \left.\frac{dK_X}{dL_X}\right|_{KMEL} > \left.\frac{dK_X}{dL_X}\right|_{DEL} > \left.\frac{dK_X}{dL_X}\right|_{LMEL} > 0$$

Finally, by comparing the elasticities of $KMEL$ and $LMEL$ with the elasticities of rays from the two origins of the Edgeworth-Bowley box (the elasticity of a ray from 0_X, expressed in terms of K_X and L_X is 1, while the corresponding elasticity of a ray from 0_Y is $\lambda_{LX}\lambda_{KY}/\lambda_{LY}\lambda_{KX}$), the following additional relationships may be derived for *any* point G along $KMEL$ and *any* point H along $LMEL$ (not just long-run equilibrium points):

$$\text{(A11)}\quad |\lambda| \gtrless 0 : \left.\frac{dK_X}{dL_X}\right|_{0_XG} \lesseqgtr \left.\frac{dK_X}{dL_X}\right|_{KMEL} \lesseqgtr \left.\frac{dK_X}{dL_X}\right|_{0_YG}$$

$$\text{and}\quad \left.\frac{dK_X}{dL_X}\right|_{0_XH} \lesseqgtr \left.\frac{dK_X}{dL_X}\right|_{LMEL} \lesseqgtr \left.\frac{dK_X}{dL_X}\right|_{0_YH}$$

B

To examine local stability when commodity prices are variable, equation (16) is equated with the economy's aggregate supply function, written (in differential form) as a function not of relative commodity prices, but of the allocations of factors to each sector:

$$\text{(A12)}\quad \hat{X} - \hat{Y} = \theta_{LX}\hat{L}_X + \theta_{KX}\hat{K}_X - \theta_{LY}\hat{L}_Y - \theta_{KY}\hat{K}_Y$$

Using (A2) and (10) to eliminate $\hat{L}_Y$ and $\hat{K}_Y$, this becomes

$$\text{(A13)}\quad \hat{X} - \hat{Y} = B_L\hat{L}_X + B_K\hat{K}_X$$

where $B_L = \theta_{LX} + \theta_{LY}\dfrac{\lambda_{LX}}{\lambda_{LY}}$

and $B_K = \theta_{KX} + \theta_{KY}\dfrac{\lambda_{KX}}{\lambda_{KY}}$

Combining (16) and (A13) with (A3) and (A5) yields differential expressions for the two factor-market equilibrium loci which take into account the endogenous changes in relative commodity prices:

$$\text{(A14)}\quad KMEL: \; -A_4\hat{K}_X + B_4\hat{L}_X = 0$$

where $A_4 = \dfrac{B_K}{\sigma_D} + \dfrac{\theta_{LX}}{\sigma_X} + \dfrac{\theta_{LY}}{\sigma_Y}\dfrac{\lambda_{KX}}{\lambda_{KY}}$

and $B_4 = -\dfrac{B_L}{\sigma_D} + \dfrac{\theta_{LX}}{\sigma_X} + \dfrac{\theta_{LY}}{\sigma_Y}\dfrac{\lambda_{LX}}{\lambda_{LY}}$

$$\text{(A15)}\quad LMEL: \; A_5\hat{K}_X - B_5\hat{L}_X = 0$$

where $A_5 = -\dfrac{B_K}{\sigma_D} + \dfrac{\theta_{KX}}{\sigma_X} + \dfrac{\theta_{KY}}{\sigma_Y}\dfrac{\lambda_{KX}}{\lambda_{KY}}$

and $B_5 = \dfrac{B_L}{\sigma_D} + \dfrac{\theta_{KX}}{\sigma_X} + \dfrac{\theta_{KY}}{\sigma_Y}\dfrac{\lambda_{LX}}{\lambda_{LY}}$

Taking a Taylor series approximation to the adjustment functions (13) and (14) (setting their elasticities equal to unity without loss of generality), and substituting from (A14) and (A15) gives the following matrix equation:

$$\text{(A16)}\quad \begin{bmatrix} DK_X \\ DL_X \end{bmatrix} = \begin{bmatrix} -A_4 & B_4 \\ A_5 & -B_5 \end{bmatrix} \begin{bmatrix} K_X - K_X^0 \\ L_X - L_X^0 \end{bmatrix}$$

Since the diagonal terms are negative, a necessary and sufficient condition for local stability is that the determinant of the coefficient matrix in (A16) be positive. This yields condition (17) in the text.

REFERENCES

Raveendra N. Batra, *Studies in the Pure Theory of International Trade*, London 1973.

J. N. Bhagwati, "The Theory of Immiserizing Growth: Further Applications," in Michael B. Connolly and Alexander K. Swoboda, eds., *International Trade and Money*, London 1973.

——— **and T. N. Srinivasan,** "The Theory of

Wage Differentials: Production Response and Factor Price Equalization," *J. Int. Econ.*, Feb. 1971, *1*, 19-35.

Mark Blaug, *Economic Theory in Retrospect*, 2d ed., London 1968.

A. Bosch, A. Mas-Colell, and A. Razin, "Instantaneous and Non-Instantaneous Adjustment to Equilibrium in Two-Sector Growth Models," *Metroecon.*, May/Aug. 1973, *25*, 105-18.

Avinash K. Dixit, *The Theory of Equilibrium Growth*, London 1976.

F. H. Hahn, "On Two-Sector Growth Models," *Rev. Econ. Stud.*, Oct. 1965, *32*, 339-46.

A. C. Harberger, "The Incidence of the Corporation Income Tax," *J. Polit. Econ.*, June 1962, *70*, 215-40.

K.-I. Inada, "Investment in Fixed Capital and the Stability of Growth Equilibrium," *Rev. Econ. Stud.*, Jan. 1966, *33*, 19-30.

H. G. Johnson and P. Mieszkowski, "The Effects of Unionization on the Distribution of Income: a General Equilibrium Approach," *Quart. J. Econ.*, Nov. 1970, *84*, 539-61.

R. W. Jones, "The Structure of Simple General Equilibrium Models," *J. Polit. Econ.*, Dec. 1965, *73*, 557-72.

———, (1971a) "Distortions in Factor Markets and the General Equilibrium Model of Production," *J. Polit. Econ.*, June 1971, *79*, 437-59.

———, (1971b) "A Three-Factor Model in Theory, Trade and History," in Jagdish N. Bhagwati et al., eds., *Trade, Balance of Payments and Growth: Essays in Honor of C. P. Kindleberger*, Amsterdam 1971.

———, "Income Distribution and Effective Protection in a Multicommodity Trade Model," *J. Econ. Theory*, Feb. 1975, *11*, 1-15.

M. C. Kemp, Y. Kimura and K. Okuguchi, "Monotonicity Properties of a Dynamical Version of the Heckscher-Ohlin Model of Production," *Econ. Stud. Quart.*, Dec. 1977, *28*, 249-53.

C. E. McLure, Jr., "A Diagrammatic Exposition of the Harberger Model with One Immobile Factor," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 56-82.

Stephen P. Magee, "Factor Market Distortions, Production, Distribution, and the Pure Theory of International Trade," *Quart. J. Econ.*, Nov. 1971, *85*, 623-43.

———, *International Trade and Distortions in Factor Markets*, New York 1976.

W. Mayer, (1974a) "Variable Returns to Scale in General Equilibrium Theory: a Comment," *Int. Econ. Rev.*, Feb. 1974, *15*, 225-35.

———, (1974b) "Short-run and Long-run Equilibrium for a Small Open Economy," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 955-67.

M. Mussa, "Tariffs and the Distribution of Income: the Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1191-204.

J. P. Neary, "Factor Tax Incidence in a Stationary Two-Sector Economy: Simplification and Extensions," mimeo, Nuffield College, Oxford 1976.

———, "Capital Subsidies and Employment in an Open Economy," *Oxford Econ. Pap.*, July 1978, *30*.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

H. Uzawa, "On a Two-Sector Model of Economic Growth," *Rev. Econ. Stud.*, Oct. 1961, *29*, 40-47.

A. L. Vandendorpe and A. F. Friedlaender, "Differential Incidence in the Presence of Initial Distorting Taxes," *J. Publ. Econ.*, Oct. 1976, *6*, 205-29.

Optimal Rewards for Economic Regulation

By MARTIN L. WEITZMAN*

Suppose several production units or firms must be regulated when costs and benefits are uncertain. Pollution might be a specific example, although there are many others. Given that firms must bear their own costs, the regulators want to transmit a schedule of revenues to each unit which in some expected value sense elicits an optimal response.

What makes this problem intriguing is that while benefits are typically a nonseparable function of *all* the firms' outputs, it seems realistic to require that the revenue function to be received by a given unit must depend in some well-defined way on *its* individual actions alone.

Two control modes often used in regulation are "prices" and "quantitites." These can be viewed as special cases of revenue functions. Prices are a linear function of output. Quantities might be described as a quadratic loss function of deviations from target, accompanied by a heavy-penalty weight. Although these two control modes are frequently treated as mutually exclusive regulatory strategies, it is highly unlikely that either extreme is optimal.

In the class of all objective functions, what is the best revenue schedule? This paper is devoted to formalizing the question, giving a precise answer (at least for an important special case), and analyzing the answer. Roughly speaking, in an optimal policy the center transmits to each firm a "price term" plus a weighted "quantity term," the weight depending in a well-defined way on specific features of the underlying situation. Such a result can be interpreted as providing a reasonable justification for regulation based on *both* price incentives and quantity targets.

I. The Regulatory Environment

As perhaps befits a theoretical paper, "regulation" is being analyzed at a rather high level of abstraction. The basic question is how to make simple rules which will induce firms to do what is best in an uncertain world. This issue is taken as the prototype problem of regulation, and it is modeled below.

The question why an economic activity must be regulated instead of being left to allocate itself in the market place is not treated directly. Possible reasons might range all the way from administrative or political considerations to one form or another of market failure. Prime examples of the kind of regulatory situation I have in mind are control of interdependent divisions in a large organization, and government regulation of externalities. In such situations there is no natural market for the good, and its production must be artificially controlled.

Suppose there are n firms or divisions to be regulated. Let x_i units of commodity i be produced by firm or division i. In the context of an externality, say pollution, x_i would be the level of the ith polluter's abatement program holding everything else constant. Depending on the interpretation, the various components of $x = (x_1, \ldots, x_n)$ might represent physically distinct goods or they could denote amounts of the same item produced by different production units.

The word "commodity" is being used in an abstract sense and really could pertain to just about any kind of good from pure water to military hardware. For the sake of preserving a unified notation we follow the standard convention of treating goods as desirable. Rather than talking about air

*Massachusetts Institute of Technology. On the occasion of his forthcoming 65th birthday, I would like to dedicate this paper to my friend, colleague, and teacher Evsey D. Domar. He fostered my interest in the problem analyzed here by puzzling aloud over the simultaneous presence of price and quantity directives in most planned systems. For their helpful comments, my thanks go to P. A. Diamond, M. Manove, J. M. Mirrlees, and the referee.

pollution, for example, I instead deal with its negative—clean air.

For a firm to produce output requires the outlay of a corresponding cost. An essential feature of the regulatory environment I am trying to describe is uncertainty about the exact specification of each firm's cost function. In most cases even the managers and engineers most closely associated with production would be unable to precisely specify beforehand the cheapest way of generating various hypothetical output levels. Because they are yet further removed from the production process, the regulators are likely to be vaguer still about a firm's cost function. This observation acquires additional force in a fast moving world where deception may be involved or where knowledge of particular circumstances of time and place may be required.

Generally speaking, there is no way the regulators can know beforehand exactly what it will cost to achieve a certain output level. Estimates can be made and the degree of fuzziness could be reduced by investigation and research. But it could never be eliminated completely because new sources of uncertainty are arising all the time. The true costs will only be known when production is actually underway.

In mathematical language, the regulators perceive the cost function of firm i as an estimate or approximation, written

$$C_i(x_i; \epsilon_i) \tag{1}$$

In the above formulation ϵ_i is a disturbance term, stochastic element, or random variable representing a state of the world unobserved and unknown at the present time. During the course of plan implementation, ϵ_i will eventually make itself known to firm i, and perhaps also to the regulators. But at the moment when an operational plan must be decided for the forthcoming period, the regulators' knowledge of ϵ_i can be represented only by a probability distribution.

The benefit function too is presumably discernable only tolerably well, say as

$$B(x; \delta) \tag{2}$$

with δ a vector of random variables having some probability distribution. The money value of various commodity output levels may be uncertain because it is imperfectly known or because authentic randomness (like the weather) is present.

It is assumed that C_i is strictly convex in x_i for each ϵ_i and B is strictly concave in x for each δ. All cost and benefit functions are presumed to be smoothly differentiable.

II. A Problem in Regulation

There is another important feature of cost functions that goes along with the uncertainty. Not only are costs unknown, but it is typically difficult and expensive to find out what they are. Sometimes economists and others share an overtendency to conceptualize regulation as a process of continual fine tuning. A certain strategy is adopted, then marginal costs and marginal benefits are observed. If they are not equal, the fees, standards, or other parameters are smoothly adjusted until an optimum is obtained.

However, this is an inappropriate way of viewing the problem. In order to be given a chance to work, a regulatory strategy must be left in place for an extended period after it has been adopted. If a firm anticipates the regulations are going to change in the near future, it is not going to take very seriously compliance with them now. This does not mean that regulations, once formulated, must be immutable for all time. It is just that they must remain in force long enough to be believable.

Another, perhaps more serious, reason that the fine tuning model may be irrelevant is that most production activity involves investment. The investment may be in research, development, reorganization, new equipment, learning by doing, etc. True costs will not become known until the investments are actually made. Whatever its form, such investment takes time and it is largely irreversible. Once made, it cannot be easily or costlessly taken back, nor can the knowledge gained be effortlessly transferred to other situations. This means that there are costs to adjusting regulations, and they are likely to be substantial.

A basic principle of regulation is that the

regulators are forced to make decisions in an uncertain environment and they must live with the consequences for some time. Among these consequences is the possibility that costs borne by some firms will turn out to be higher or lower than was expected. A good regulatory strategy will take advantage of this by instituting a reward structure which automatically encourages the cheap firm to produce more and the expensive firm less. In our formulation, regulators are confined to a strategy of indirect control by judiciously selecting revenue functions in advance for each firm.

Now, in a certain sense the ideal revenue function for any firm is the entire expected benefits function, plus or minus some constant. Assuming away the game-theoretic problems having to do with bluffing, threatening, etc., a Nash-type equilibrium might conceivably emerge where each firm would have the incentive to set its marginal cost equal to its marginal benefit after all uncertainty had been eliminated and every firm knew what every other firm was doing.

The trouble with this sort of approach is that benefits are typically a nonseparable function of *all* the firms' outputs, whereas a particular firm has control only over its *own* output. It seems like a relevant abstraction to insist that a regulatory agency cannot reward or penalize a firm in what might be viewed as an arbitrary or capricious manner. Asking a firm to bear the extra risk involved in adopting a revenue schedule depending on uncertain variables not under its control may be infeasible or unacceptable. Some of the reasons for this have just been cited in downplaying the relevance of the fine tuning model. In addition, such a schedule may simply be too complicated to handle.

That revenue functions should depend only upon individual actions is a strong assumption (for example, it rules out profit-sharing incentive schemes), but I think it is appropriate to the kind of regulatory environment I have described. In this paper I take as a point of departure a scenario where firms pay their own costs and the state sets revenue functions for each firm which depend *only* on that firm's output.

III. A Formulation of the Basic Problem

A *revenue function* $R_i(x_i)$ is a schedule of monetary payments received by firm i as a function of its output. For example, if a price p_i is paid for the output of firm i, the corresponding revenue function is

$$R_i(x_i) = p_i x_i \tag{3}$$

Or, if it is the intention of the planners to set a quota $\hat{x}_i$, they might specify the following revenue function:

$$R_i(x_i) = \frac{-q_i}{2}(x_i - \hat{x}_i)^2 \tag{4}$$

where q_i is a large number.

It is important to realize that the process of profit maximization causes every revenue function to generate some output response. The response depends on the revenue function $R_i(\ )$ and the state of the world ϵ_i. For a given $R_i(\ )$ and ϵ_i, firm i will set its output x_i at that level which maximizes profits, implicitly solving the equation

$$\max_{x_i \geq 0} R_i(x_i) - C_i(x_i, \epsilon_i) \tag{5}$$

Equation (5) should not be interpreted too literally as saying that the firm knows the exact value of ϵ_i with certainty (at the same time the regulator knows only the probability distribution of ϵ_i). In the scenario I have in mind, when a revenue function is instituted for a sufficiently long period the firm will eventually grope its way to a profit-maximizing output, presumably by trial and error testing of the relevant alternatives. This is quite a different interpretation from having the cost function known a priori.

Without any significant loss of generality in the problem to be posed, we limit attention to revenue functions which generate *unique* output responses. That is, the solution of (5) is some response *function*

$$x_i = G_i(R_i(\ ), \epsilon_i) \tag{6}$$

satisfying for all possible ϵ_i the condition

$$R_i(G_i(R_i(\ ), \epsilon_i)) - C_i(G_i(R_i(\ ), \epsilon_i); \epsilon_i) = \max_{x_i \geq 0} R_i(x_i) - C_i(x_i; \epsilon_i) \tag{7}$$

Note that changing a revenue function by adding or subtracting any constant cash payment does not alter the corresponding response (aside from the issue of setting such a low payment that the firm is forced out of business altogether). At least in a rough way, this might be interpreted as providing some justification for studying the allocative effects of a revenue function apart from the distributive consequences.

In the framework adopted here, the planners are at a point where as much information as is feasible to gather has already been obtained. An operational plan must now be decided on the basis of the available current knowledge, summarized by (1) and (2). Because it will force long-term resource commitments (like capital investments), any incentive scheme has serious consequences which continue for some time and cannot easily be reversed. This is an essential feature of the regulatory environment very prominent, for example, in the case of pollution. A regulatory agency must resign itself to naming in advance revenue functions $\{R_i(\ )\}$ and living with the outcome even though it does not presently know the values of $\{\epsilon_i\}$ or δ.

Through the output response (6) which they induce, reward functions $\{R_i(\ )\}$ yield the expected differences in benefits and costs

$$(8)\quad \Phi(\{R_i(\ )\}) \equiv \underset{\{\epsilon_i\},\delta}{E}\Big[B(\{G_i(R_i(\ ),\epsilon_i)\};\delta) - \sum_{i=1}^{n} C_i(G_i(R_i(\ ),\epsilon_i);\epsilon_i)\Big]$$

A set of *optimal* revenues $\{R_i^*(\ )\}$ is any collection of functions which maximize (8). In other words, via the output response generated by them, optimal revenue functions maximize expected benefits minus costs. This can formally be written[1]

$$(9)\quad \Phi(\{R_i^*(\ )\}) = \max_{\{R_i(\ )\}} \Phi(\{R_i(\ )\})$$

[1] The maximization is over the class of all possible reward functions yielding response functions. It is not difficult to prescribe conditions which ensure the existence of a solution to (9).

The above problem shares certain features of the more general structure analyzed in the theory of teams. Indeed, one of the more significant results of team theory will be used in proving the basic theorem of this paper.

Note that (9) is easy to solve when the benefit function is additively separable in the output of each firm. Then the optimal revenue function for a firm is just *its* part of expected benefits. The interesting case is where the benefit function is not separable.

IV. Optimal Revenue Functions

The remainder of this paper is devoted to characterizing the form of an optimal revenue function and explaining its dependence on various factors. Under the most general circumstances this appears to be a very intricate task. Fortunately a complete characterization is possible for an important special case.

Optimal revenue functions generate a range of output responses as the uncertainty varies. From now on it will be assumed that within this output range marginal costs and marginal benefits can be accurately approximated by linear forms.

A linear approximation might be rationalized on one of two grounds. The amount of uncertainty could be small enough to keep the range of output responses sufficiently limited to justify a first-order approximation. Or, it might just happen that total cost and benefit functions are almost quadratic to begin with. At any rate, the possibility of sharply characterizing an optimal solution makes the linear case a natural preliminary to any more general analysis.

Consider for a moment the problem of finding an optimal set of quotas or targets $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$. The optimal quota maximizes expected benefits minus expected costs, so that

$$(10)\quad \underset{\{\epsilon_i\},\delta}{E}\left[B(\hat{x},\delta) - \sum_{i=1}^{n} C_i(\hat{x}_i;\epsilon_i)\right] = \max_{x} \underset{\{\epsilon_i\},\delta}{E}\left[B(x;\delta) - \sum_{i=1}^{n} C_i(x_i;\epsilon_i)\right]$$

Presuming it is interior, the solution of (10) satisfies the first-order condition

$$(11)\quad p_i \equiv \underset{\delta}{E}\, B^i(\hat{x};\delta) = \underset{\epsilon_i}{E}\, C_i'(\hat{x}_i;\epsilon_i) \qquad i = 1,\dots,n$$

where p_i is the expected marginal benefit equals marginal cost of the ith commodity evaluated at the optimal quota.[2]

Under the linearity assumption, the marginal cost of the ith producer can be written

$$(12)\quad C_i'(x_i;\epsilon_i) = p_i + \gamma_i(x_i - \hat{x}_i) + \epsilon_i \qquad i = 1,\dots,n$$

while the marginal benefit of commodity i can be expressed as

$$(13)\quad B^i(x;\delta) = p_i - \sum_{j=1}^{n} \beta_{ij}(x_j - \hat{x}_j) + \delta_i \qquad i = 1,\dots,n$$

where, without loss of generality,

$$(14)\quad E\delta_i = E\epsilon_i = 0 \qquad i = 1,\dots,n$$

The various δ_i are just components of δ.

In order to obtain sharp results, a further regularity assumption on the probability distributions is needed. The conditional expectation of ϵ_j given ϵ_i is presumed proportional to ϵ_i. Likewise for the expected value of δ_i conditional on ϵ_i. That is,

$$(15)\quad E\epsilon_j/\epsilon_i = \theta_{ji}\epsilon_i \qquad i = 1,\dots,n$$
$$E\delta_i/\epsilon_i = \eta_i\epsilon_i \qquad j = 1,\dots,n$$

for some coefficients $\{\theta_{ji}\}$, $\{\eta_i\}$. Naturally $\theta_{ii} = 1$.

Condition (15) could be justified as a first-order approximation holding for small uncertainties. It would also be a consequence of a joint normal distribution in $\{\epsilon_i\}$ and $\{\delta_i\}$. All independent probability distributions ($\theta_{ji} = 0$, $j \neq i$) automatically satisfy (15).

Note that

$$\theta_{ji} = \sigma_{ij}^2/\sigma_{ii}^2,\ \eta_i = \sigma_{i0}^2/\sigma_{ii}^2$$

where $\quad \sigma_{ij}^2 = E\epsilon_i\epsilon_j,\ \sigma_{i0}^2 = E\epsilon_i\delta_i$

[2]All the $\{p_i\}$ are identical when the various commodities represent the same item produced by different production units.

The basic result of the present paper is summarized by the following:

THEOREM: *Under the assumptions* (12)–(15), *the optimal reward function for unit i can be expressed in the form*

$$(16)\quad R_i^*(x_i) = p_i x_i - \frac{q_i}{2}(x_i - \hat{x}_i)^2 \pm \textit{constant}$$

The $\{q_i\}$ *are coefficients satisfying*[3] *the equations (linear in* $\{1/q_i + \gamma_i\}$)

$$(17)\quad \sum_{j=1}^{n} \frac{\beta_{ij}\theta_{ji}}{q_j + \gamma_j} + \eta_i = 1 - \frac{\gamma_i}{q_i + \gamma_i} \qquad i = 1,\dots,n$$

V. Analysis of an Optimal Reward

Equation (16) means that aside from the arbitrary constant, an optimal reward function can be decomposed into two components.

The first term

$$(18)\qquad p_i x_i$$

is the traditional price signal. If p_i accurately represented the marginal benefit of commodity i, using (18) as a reward function would automatically induce firm i to produce at that output level where marginal benefit equals marginal cost. The apparent guarantee of social efficiency is what makes the use of prices as a regulatory device so attractive to the economist. Unfortunately for this idea, the marginal benefit of commodity i cannot be reduced to a single number which is precisely known beforehand.

[3]It is assumed that (17) has a solution and that $\{q_i + \gamma_i\}$ are positive. The latter condition is needed to guarantee that the problem of maximizing revenues (16) minus costs has a meaningful solution for each firm. Although it seems hard to prove a very broad sufficiency theorem, playing with lots of examples has convinced me that the condition holds for most cases of economic interest. A sufficiency theorem can be proved when firms are close to being symmetric with each other or as the uncertainties are not too far from being independently distributed.

The second component of (16),

$$\frac{-q_i}{2}(x_i - \hat{x}_i)^2 \tag{19}$$

is a quadratic penalty for departures from the target $\hat{x}_i$. Were $\hat{x}_i$ in fact the socially optimal output of commodity i, the center could do no better than transmitting (19) as a revenue function, with q_i arbitrarily large (which is equivalent to setting $\hat{x}_i$ as a standard). The seeming ability to directly fix economic activity at the socially desirable level is what makes the quota appealing as a regulatory device, especially to the general public. Alas, the regulators don't know exactly what output levels are socially optimal to begin with.

The basic result of this paper argues that in economic planning situations prices and quotas are not redundant or inconsistent messages. In fact, a "mixed" price-quota system is the optimal reward. The coefficient q_i determines the composition of the mix. With $q_i = 0$, (16) becomes a pure price signal, when $q_i = \infty$, (16) is made into a complete quota system.[4]

All this strongly suggests that a regulatory strategy based on *both* price incentives and quantity targets, far from being a contradiction, is actually optimal in a world of uncertainty. Such a principle has been intuitively sensed, I believe, by practical planners. Of course the revenue function is usually not formalized as it is in (16). Instead there is typically some vaguely ambiguous policy of rewarding output while simultaneously discouraging deviations from a target. Taking advantage of the theorist's inherent right of simplification, I would suggest that (16) is not a bad translation of such a policy.

A result like (16) provides at least a partial resolution of the environmental economics debate between the use of effluent charges and the use of effluent standards. It also offers some justification for mixed price and quantity controls within a large, divisionalized organization.

[4]The comparative advantage of these two extreme regulatory modes was analyzed in my 1974 article.

Note that the optimal reward function does not promise social optimality or efficiency *ex post*, after $\{\epsilon_i\}$ and $\{\delta_i\}$ take on specific values. The concept of *ex post* social optimality is too strong to require, given the informational constraints being imposed. The relevant issue is which reward function comes closest to inducing a social optimum in some average sense.

A simple description explains how $\{\hat{x}_i\}$ and $\{p_i\}$ are determined. They are just the optimal outputs and their marginal values obtained when the center, suppressing all uncertainty, maximizes the difference between the "representative" benefit function

$$b(x) \equiv \underset{\delta}{E}\, B(x;\delta)$$

and the sum of "representative" cost functions

$$c_i(x_i) \equiv \underset{\epsilon_i}{E}\, C_i(x_i;\epsilon_i)$$

The determination of the penalty weights $\{q_i\}$ is slightly more complicated to explain. Differentiating (16) and setting the resulting expression equal to (12) yields the response function

$$x_i(\epsilon_i) = \hat{x}_i - \frac{\epsilon_i}{q_i + \gamma_i} \tag{20}$$

Suppose $\epsilon_i = 1$. This lowers the output of x_i by $1/(q_i + \gamma_i)$ units, causing a net increase of

$$1 - \frac{\gamma_i}{q_i + \gamma_i} \tag{21}$$

dollars in the marginal cost of firm i. By (15), δ_i is expected to be η_i dollars above average while ϵ_j is expected to be θ_{ji} dollars above its mean. From (15) and (20), firm j can be expected to curtail output by $\theta_{ji}/(q_j + \gamma_j)$ units. Using (13), the marginal benefit of commodity i is expected to increase by

$$\eta_i + \sum_{j=1}^{n} \frac{\beta_{ji}\theta_{ji}}{q_j + \gamma_j} \tag{22}$$

dollars. Equating the increase in marginal

costs (21) with the expected increase in marginal benefits (22) yields condition (17).

On what things do the coefficients $\{q_i\}$ depend? To strengthen our intuitive feeling for the meaning of equation (17), let us turn first to a special case which can serve as a point of departure.

Suppost that all the uncertainties are independent, so that

$$(23) \qquad \theta_{ji} = 0 \qquad \text{for } j \neq i$$
$$\eta_i = 0$$

In this case equation (17) reduces to

$$(24) \qquad q_i = \beta_{ii}$$

Under (23) the optimal penalty coefficient for a commodity is just the curvature of the benefit function in that commodity. With independent uncertainties firm i should be given as a reward that part of expected benefits which remains as a function of *its* output when the outputs of all other firms $j(j \neq i)$ have been parametrically fixed at $\hat{x}_j$. This interpretation comes from examining (13), (16), and (24)

The greater the curvature in benefits, the more significant is the weight of the quantity term (19) in the reward function mix (16) If marginal benefits decrease rapidly around the optimal quota, there is a high degree of risk aversion and the center cannot afford being even slightly off the mark. Relying too much on the price mode is risky because a miscalculation results in under or overshooting the target, with detrimental consequences. In such a situation the quantity mode scores a lot of points because a high premium is put on the rigid output controllability which only it can provide under uncertainty.

On the other hand, the weight of the quantity term is lessened when benefits are closer to being linear. In that case it would be foolish to place too much emphasis on targets. Since expected marginal social benefit is approximately constant over some range, a superior policy comes closer to naming it as a price and letting the producer find the optimal output level himself after eliminating the uncertainty from costs.

Returning to the more general case, when (23) does not hold q_i will tend to exceed β_{ii}. Generally speaking, the penalty coefficients $\{q_i\}$ become more significant as the interdependence coefficients $\{\theta_{ij}\}$ and $\{\eta_i\}$ increase.

The reason for this is easy to understand. A positive θ_{ji} means that when the marginal costs of firm i are low, so are those of firm j. Whenever firm i is increasing output because its costs are low, firm j is doing likewise. Compared with a situation of independent marginal costs, more damping should be introduced; this would stabilize welfare decreasing over and underreactions in aggregate output responses.

Something analogous happens in the case of positive η_i. Producers will cut back output for higher marginal costs, but this cutback should be dampened when there tends to be a simultaneous increase in marginal benefits. In such situations a greater weight for the quantity mode is appropriate because that mode has better properties as a stabilizer. The story is the other way round when η_i is negative.

The coefficient β_{ij} is a measure of the degree of complementarity between commodities i and j When it is higher, more stabilizing is desirable to keep commodities i and j closer to their appropriate proportions. When it is lower, less weight is needed on individual quantity terms because greater substitutability is possible.

An instructive illustration of what the quantity weights depend on is provided by the special regularized case of perfect symmetry:

$$\gamma_i = \gamma \qquad \eta_i = \eta$$
$$\theta_{ii} = 1 \qquad \beta_{ii} = \beta$$
$$\theta_{ij} = \rho \quad i \neq j \qquad \beta_{ij} = \mu\beta \quad i \neq j$$

In this case (17) yields

$$q_i = q = \frac{\eta\gamma + \beta + (n-1)\mu\beta\rho}{1-\eta}$$

The quantity weight q increases in β, ρ, η, and μ, verifying our previous discussions.

VI. Proof of the Main Proposition

Consider the problem of finding a set of optimal response functions $\{x_i(\epsilon_i)\}$ in an information structure where firm i observes only ϵ_i and controls only x_i. The "objective function" is the expected difference between benefits and costs. The problem is to maximize over $\{x_i(\epsilon_i)\}$ the function

$$(25)\quad \psi(\{x_i(\epsilon_i)\}) = \underset{\{\epsilon_i\},\delta}{E}\left[B(\{x_i(\epsilon_i)\};\delta) - \sum_{i=1}^{n} C_i(x_i(\epsilon_i);\epsilon_i)\right]$$

Since by assumption (6) any revenue function generates a response function, the solution to the above problem yields at least as high a value of the objective function as the solution to problem (9).

Now it turns out that finding optimal response functions in the present framework is an example of a classical problem in the theory of teams, whose solution[5] is given by (20) with the definition (17). Because the general result is typically presented in a somewhat different framework and may be difficult to follow, I will sketch a proof for the case treated here.

For simplicity, assume discrete distributions. Let ϵ_{ii} be a value which ϵ_i takes on with positive probability. Let $x_{ii} = x_i(\epsilon_{ii})$ be the output response of firm i when $\epsilon_i = \epsilon_{ii}$. It is not difficult to show that $\psi(\ \ )$ in (25) is a concave differentiable function of the variables $\{x_{ii}\}$ over which it is being maximized (this derives essentially from the concavity-differentiability of the benefit minus cost function and the concavity-differentiability preserving properties of an expected value operator). Hence the appropriate first-order conditions are necessary and sufficient for an optimum.

[5]See Jacob Marshak and Roy Radner, Theorem 5, p. 168. Matching up my notation with theirs is a bit messy, and it seemed better to omit the details, which the interested reader should be able to supply. I am indebted to Kenneth J. Arrow for pointing out to me that my characterization of optimal response functions is really a special case of Radner's result.

From (12), the marginal expected cost of x_{ii} given $\epsilon_i = \epsilon_{ii}$ is

$$(26)\quad E\left[\frac{\partial C_i}{\partial x_{ii}}\bigg|\epsilon_{ii}\right] = p_i + \gamma_i(x_{ii} - \hat{x}_i) + \epsilon_{ii}$$

From (13), the marginal expected benefit of x_{ii} given $\epsilon_i = \epsilon_{ii}$ is

$$(27)\quad E\left[\frac{\partial B}{\partial x_{ii}}\bigg|\epsilon_{ii}\right] = p_i - \sum_{j=1}^{n} \beta_{ji}(E[x_j \mid \epsilon_{ii}] - \hat{x}_j) + E[\delta_i \mid \epsilon_{ii}]$$

Equating (27) with (26) yields the first-order condition

$$(28)\quad -\sum_{j=1}^{n} \beta_{ji}(E[x_j \mid \epsilon_{ii}] - \hat{x}_j) + E[\delta_i \mid \epsilon_{ii}] = \gamma_i(x_{ii} - \hat{x}_i) + \epsilon_{ii}$$

We must verify that the solution proposed in (20) satisfies (28). Therefore, for x_j substitute

$$(29)\quad x_j = \hat{x}_j - \frac{\epsilon_j}{q_j + \gamma_j}$$

and for x_{ii} substitute

$$(30)\quad x_{ii} = \hat{x}_i - \frac{\epsilon_{ii}}{q_i + \gamma_i}$$

Plugging (29) and (30) into (28) yields

$$\sum_{j=1}^{n} \frac{\beta_{ji}E[\epsilon_j \mid \epsilon_{ii}]}{q_j + \gamma_j} + E[\delta_i \mid \epsilon_{ii}] = -\frac{\gamma_i\epsilon_{ii}}{q_i + \gamma_i} + \epsilon_{ii}$$

Using (15), the above expression becomes

$$(31)\quad \epsilon_{ii} \sum_{j=1}^{n}\left(\frac{\beta_{ji}\theta_{ji}}{q_j + \gamma_j} + \eta_i\right) = \epsilon_{ii}\left(1 - \frac{\gamma_i}{q_i + \gamma_i}\right)$$

Equation (31) will hold for all possible ϵ_{ii} if the $\{q_i\}$ are defined by (17). Thus, expression (20) is indeed the optimal response function.

The remainder of the proof consists of verifying that under cost assumption (12),

the revenue function (16) generates[6] the response function (20).

[6]Recall I am assuming the $\{q_i + \gamma_i\}$ satisfying (17) to be positive. If the solution to (17) yields a negative value of $q_i + \gamma_i$ for some i, there will still exist an optimal response function (in a team theory sense), given by (20). Unfortunately, there is no way to induce firm i to follow this rule by naming a corresponding revenue function The problem is that with a negative $q_i + \gamma_i$, (20) dictates that the firm should produce *more* when its costs are *higher* This kind of seemingly perverse behavior might be optimal if, for example, whenever the costs of firm i are high, the costs of other firms are much higher still. But no revenue function can elicit such a perverse response from firm i The regulators would have to rely on moral suasion or some other means. The team theory approach would make no distinction between positive and negative $q_i + \gamma_i$ But the reliance on revenue functions to elicit proper behavior requires that $q_i + \gamma_i$ be positive for all firms, at least for the theorem proved here

REFERENCES

Jacob Marshak and Roy Radner, *Economic Theory of Teams*, New Haven 1972.

M. L. Weitzman, "Prices vs. Quantities," *Rev. Econ. Stud.*, Oct. 1974, *41*, 477 91.

Security Price Changes and Transaction Volumes: Additional Evidence

By MARK HANNA*

In a recent issue of this *Review*, Thomas W. Epps reported that volume relative to price change was generally greater on upticks than downticks for successive bond transactions (i.e., that $V^+/P^+ > V^-/P^-$, where V^+ and V^- are volume on upticks and downticks, respectively, and P^+ and P^- are price changes on upticks and downticks, respectively). Although his test incorporated a large number of observations for each of twenty bonds, it was limited to a single month whose uniqueness may have seriously biased his findings. With respect to price changes, the month used was unique. To test for possible distortions stemming therefrom, this comment incorporates the results of replicating his study using a month whose primary characteristic is diametrically opposed to that Epps used. In addition, components of Epps' ratio (volume/price change) are examined separately for the possibility of overwhelming dominance by one component as the explanation of differences in ratio behavior on upticks vis-à-vis downticks.

Epps' results are summarized under "January" in Table 1: sixteen of the twenty bonds in his sample supported his theoretical construct. His thesis is supported by a negative "z" statistic using a Wilcoxon test. Moreover, Fisher's test of combined results indicates that the probability of obtaining such negative outcomes by chance was less than .031. I have precisely replicated his study using the same month, January 1971 (*JAN*), and my findings confirm his *JAN* findings.[1]

Epps does not comment upon the month he uses other than to identify it. But *JAN* saw one of the sharpest price upswings of recent history. A decline of fifty basis points in the interest rate of new issues of Aaa corporate bonds was the fourth greatest monthly decline of the last decade or so (1964-76).[2] Expressed in relative terms (change in basis points/ interest rate), it was the fifth greatest monthly decline in interest rates. Clearly, the sample month was not a typical month. *If* sharply rising prices bias Epps' results, his study would be subject to strong bias.

To assess the possibility of bias from this cause, I have replicated Epps' study in a month of sharply declining bond prices. May 1971 (*MAY*) saw a rise of forty-nine basis points in yield, the second greatest rise of the 1964-76 period. The greatest monthly rise was fifty-nine basis points in July 1974. However, proportionately *MAY* saw the greatest rise in interest rates.[3] *MAY* is an especially appropriate month for testing whether Epps' results were influenced by price trend - not only because it had the opposite price trend from *JAN*, but also because of its proximity to *JAN*. Proximity is desirable because it tends to hold the respective maturities of the component bonds constant.[4]

The findings for *MAY* are also presented in Table 1. Sixteen of the bonds supported Epps' contention with respect to sign and four contradicted it. In number, this find-

*Professor of finance, University of Georgia. I wish to thank Charles C. Kidd for his valuable assistance.

[1]The data source was also the same, namely F. E. Fitch.

[2]See the *Treasury Bulletin*, pp. 81-82.

[3]In *MAY* the twenty bonds did not evince the uniformity of trend found in *JAN*. While in *JAN* all the bonds were up in price, in *MAY* three of the bonds countered the general downtrend. Each bond in *JAN* saw more up days than down days; in *MAY*, while the bulk had more down days than up days, there were three with more up days. Still, overall, *MAY* unquestionably qualifies as a month of substantial downtrend in prices.

[4]Not all the bonds are long term. Fourteen mature beyond 1992 but six mature(d) prior to 1978. Hence, the use of an index of long-term yields to identify large price swings is not entirely appropriate.

TABLE 1—z STATISTICS FROM WILCOXON TESTS OF THE DIFFERENCE IN CENTERING BETWEEN THE DISTRIBUTION OF V^+/P^+ AND V^-/P^-

Bond	Wilcoxon z Statistics[a] JANUARY	MAY
Chrysler Corp. 8 7/8s95	.550	1.612
Chase Manhattan 4 7/8cv93	− 427	−1.379
Commonwealth Ed. 8 3/4s75	1.852	−1.135
Dow Chemical 8 7/8s2000	.486	−2.164
Ford Motor Co. 8 1/4s74	− .647	− .166
Ford Motor Credit 8 3/4s75	− .632	− 257
G. M. A. C. 8 3/4s77	−1.917	− .880
Standard Oil of N. J 6s97	− .158	−1.181
J. C. Penney 8 7/8s95	553	− .143
Nat'l. Cash Register 6cv95	−1 885	− .665
N J. Bell T. 9 35s10	− .566	− 831
New Eng T & T 8 5/8s09	− 790	.296
Sears Roebuck 8 1/8s76	− 766	−. 868
S. Central Bell 8 1/4s04	− 277	−1 012
Santa Fe Ind 6 1/4cv98	−1.272	328
Signal Companies 8.85s94	− 312	−2 000
Standard Oil of Ind. 6s98	− .383	− .809
Swestern Bell T 8 3/4s07	− 184	− .547
A T. & T 8 3/4s2000	−2 513	−1 021
Union Oil of Cal. 8 1/4s76	− .178	.301

Sources. z statistics for *JAN* were taken from Epps (1975, Table 1, p 595). Original data for *MAY* were taken from Fitch, as were data for *JAN* *MAY* z statistics were calculated by the author

[a]Noncumulative volume specification is used which is explained in fn 8

ng coincides exactly with Epps' for *JAN*. The cast of the four with wrong signs, however, changes, with only Chrysler having wrong signs in both studies.) While the total number of negative z statistics was the same for both months, their distribution indicated even more statistical significance in support of Epps in *MAY* (Fisher's test *alpha* = .0142) than in *JAN* (*alpha* = .031). Therefore, one must conclude that the choice of *JAN* did not bias Epps' findings.

The possible dominance by one of the two components in Epps' ratio is examined next. In the discussion that follows:

$$R^+ = V^+/P^+, \text{ while } R^- = V^-/P^-$$
$$\bar{R}^+ = \bar{V}^+/\bar{P}^+, \text{ while } \bar{R}^- = \bar{V}^-/\bar{P}^-$$

with a bar signifying the mean.

While Epps warns that the Wilcoxon test is based on the whole distribution of R^+ vis-à-vis R^- for a particular bond, the use of simple averages may be fruitful in gaining insights into the relative importance of the components of the ratios as determinants of the z statistic's sign. If the average is a suitable proxy for the distribution (a likelihood strengthened by the fact that in the Wilcoxon test absolute size of observed ratios is subjected to a relative size (ranking) transposition), then a bond whose $\bar{R}^+ > \bar{R}^-$ would be expected to evince a negative z statistic based on a Wilcoxon whole distribution test.[5] Likewise, if $\bar{R}^+ < \bar{R}^-$, we would expect a positive z. And, in fact, when *JAN* and *MAY* are aggregated, the $\bar{R}^+$, $\bar{R}^-$ relationship for the individual bonds was consistent with the respective actual z by a 33 7 margin (82.5 percent agreement). Since actual z statistics are predominantly negative (32–8), such a large

[5]Clearly $\bar{R}^+ = \bar{V}^+/\bar{P}^+ \neq (\overline{V^+/P^+})$ but it is desirable to use the former for $\bar{R}^+$ since V^+ is to be compared to $\bar{V}^-$ (and $\bar{P}^+$ to $\bar{P}^-$) in the next step

margin of sign consistency means that for the most part the $\bar{R}^+$, $\bar{R}^-$ relationship is generally consistent with a *negative* z. And, in fact, this is the case by a 25–15 margin. A distribution as extreme as this one is hardly likely to occur by chance (less than 7.7 percent of the time for a binomial distribution where the expected signs were distributed 50–50). Thus, averages reflect the same negative z tendency as found using the whole distributions. Averages serve as suitable, though hardly perfect, proxies for the whole distributions. In a parallel manner, averages will now be used in analyzing the components of the $\bar{R}$s.

For a particular bond where $\bar{V}^+ > \bar{V}^-$, we would expect a negative z: for the magnitudes of $\bar{R}^+$ and $\bar{R}^-$ would vary with their numerators. Likewise where $\bar{P}^+ < \bar{P}^-$ for the ratios would vary inversely with their denominators. Examining these component relationships separately, we find the $\bar{V}^+$, $\bar{V}^-$ relationship consistent with negative z statistics by a margin of 25–15 (significant at *alpha* .077 while the $\bar{P}^+$, $\bar{P}^-$ relationship is only consistent by a 22–18 margin (not even significant at the .31 level).

The dominating role of the V relationship over that of P is emphasized by noting the subperiods separately. In *MAY* $\bar{P}^+$ is exceeded by $\bar{P}^-$ by a 14–6 margin (significant at $\alpha = .0577$), strongly consistent with a negative z. But this consistency is explained by the fact that prices trended down in *MAY*. In such periods we expect the size of average upticks to be exceeded by the average for downticks.[6] In *JAN* the relationship reverses as prices rise: $\bar{P}^+ > \bar{P}^-$ by a 12–8 margin, on balance consistent with a *positive* z. Clearly, any importance of the P relationship as a determinant of a negative z tendency depends on market trend and not on a general tendency inherent in the P^+, P^- relationship.

The V relationship, on the other hand, remains consistent with a negative z regardless of price trend; by a 13–7 margin in *JAN* and 12–8 in *MAY*. Thus, the V relationship is the dominant determinant of the negative z statistic; $\bar{V}^+$ tends to exceed $\bar{V}^-$ in declining as well as rising price periods.[7]

In conclusion, rather than there being bias in favor of Epps' thesis because of the sharp uptrend in price during January, the bias from price trend was, in fact, in the opposite direction. The tendency for the size of upticks to exceed that of downticks (P^+ to exceed P^-) in uptrends, and therefore for the price behavior tendency to cause V^+/P^+ to be less than V^-/P^-, is a bias that must be overcome in order for Epps' thesis to be fulfilled. But that Epps' thesis was generally supported even in an uptrend month, where there is a bias against such a finding due to price trend, emphasizes that the chief determinant of the ratio relationship is volume behavior. For the tendency of volume to pick up on upticks and dry up on downticks in both rising and falling markets was sufficiently strong to overcome a strong contrary bias stemming from price in the month of rising prices.[8] This dominance of the influence of volume on the ratio may call for modification in Epps' model, but otherwise this extension of Epps' empirical work supports his findings.

[6] In downtrends, downticks tend to exceed upticks in number and size. For the typical bond, in *JAN* upticks prevailed (53.4 percent), in *MAY*, downticks (52.3 percent).

[7] Bond-by-bond details of the components' "consistency" tests are available upon request.

[8] In addition to the above tests, daily volume and changes in price between days instead of between transactions were examined to see if the same tendency existed from day to day as Epps found from trade to trade. Epps' noncumulative volume model was used, meaning trades within a day of price change are cumulated but additional days at the same closing price are ignored. In *JAN* ten bonds were consistent with Epps' thesis, nine contrary, and one with a z of zero. In *MAY* eleven were consistent (zs were negative), eight contrary, and one had too few observations to be tested meaningfully. These distributions differ hardly at all from chance ones. Moreover, there is practically no correlation within each month between the z based on continuous trades and that based on daily intervals (−.15 and −.10, respectively). Clearly, the tendency Epps found concerning continuous trades within a single month is not confirmed using daily (aggregated) data. However, limitation of observations to the number of trading days per month inherent in this test precludes drawing conclusions about longer-run tendencies. Epps, in another paper (1977), found that daily price-volume data for common stocks exhibit his predicted characteristics.

REFERENCES

T. W. Epps, "Security Price Changes and Transaction Volumes: Theory and Evidence," *Amer. Econ. Rev.*, Sept. 1975, *65*, 586–97.

———, "Security Price Changes and Transaction Volumes: Some Additional Evidence," *J. Finan. Quant. Anal.*, Mar. 1977, *12*, 141–46.

F. E. Fitch, *Bond Sales on the New York Stock Exchange, A Daily Market Publication*, New York, Jan; May 1971.

U.S. Treasury Department, *Treasury Bull.*, Mar. 1977.

Security Price Changes and Transaction Volumes: Comment

By MEIR I. SCHNELLER*

In a recent issue of this *Review*, Thomas W. Epps attempted to provide a theoretical framework which would explain "a familiar Wall Street adage ... that transaction volume (the number of shares traded) tends to be relatively high in bull markets and low in bear markets ... such a theoretical framework can be constructed from a fairly broad set of portfolio-selection models..." (p. 586). The portfolio-selection model which was used by Epps observes a market divided between bullish and bearish investors, all of whom are mean-variance decision makers. In this note I show that Epps' assumptions are inconsistent with those of the mean-variance model. Furthermore, it is shown that even if we accept the mean-variance framework as a working assumption, the results claimed by Epps are incorrect.

It has long been maintained[1] that the mean-variance portfolio-selection model is valid only if one of the following conditions is satisfied: 1) that the distributions under consideration are normal; or 2) that investors are characterized by quadratic utility functions. Epps' assumptions contradict both of these conditions: his prospects are lognormally distributed while the utility functions that he assigns to his investors cannot be quadratic. The reason is that the quadratic utility function does not satisfy Epps' requirement of constant "extent of risk aversion" (denoted by β)[2] which implies linear indifference curves in the mean-variance space. I shall now show that even if we accept the mean-variance framework as a working assumption[3,4] Epps' results do not follow.

The main thrust of Epps' analysis lies in the segmentation of the market into two parts: a bullish segment (indexed by T) and a bearish segment (indexed by U). Based on his assumptions Epps derives the demand function of the two segments:[5]

$$P = \Psi_T - \Phi_T a_T \tag{19}$$

$$P = \Psi_U - \Phi_U a_U \tag{20}$$

where we derive from his (18)

$$\Psi = \left(\sum_{j=1}^{J} \frac{1}{\gamma_j}\right)^{-1} \sum_{j=1}^{J} \frac{\alpha_j p_j}{\gamma_j}$$

$$\Phi = \left(\sum_{j=1}^{J} \frac{1}{\gamma_j}\right)^{-1}$$

and

$$\alpha_k = 1 + \frac{S_k}{p_k} \sum_{i \neq k} \frac{r_{ik}(p_i - P_i)}{r_{kk} S_i} \tag{14}$$

$$\gamma_k = S_k^2 \beta \frac{r}{r_{kk}} \tag{15}$$

and P_i is the current market price of security i. For the definitions of S_k, p_i, r_{ik}, and r_{kk}, the reader is referred to Epps. When a bullish information is observed by the bullish investors but ignored by the bearish

*Lecturer, The Jerusalem School of Business, The Hebrew University.

[1] See James Tobin.

[2] β is defined by Epps as $(\partial u/\partial var)/(\partial u/\partial exp)$ where *var* and *exp*, are the variance and expected value of end of period wealth.

[3] Paul Samuelson has shown that where the distributions under consideration are compact, the mean-variance criterion yields a portfolio which is a good approximation of the actual optimal portfolio even if both conditions for the mean-variance model are violated

[4] I also adopt the rest of Epps' assumptions despite the fact that the reasonability of some of them are quite questionable, for example, assumption 9 which assigns a high degree of risk aversion to bullish investors, and a low degree of risk aversion to bearish investors

[5] The reader should interpret these equations carefully, whereas the index k in equations (14) and (15) refers to the kth security, the index j in equation (18) refers to investor j who observes security k.

investors, the initial equilibrium price $P_0 = \Psi_U - \Phi_U a_{U0}$ changes according to Epps to $P_1 = \Psi_U - (\Phi_U a_{U0} - V^+)$ and as a result:

$$\frac{V^+}{\Delta P^+} = \frac{1}{\Phi_U}$$

Likewise, the arrival of bearish information ignored by bullish investors will bring about the equality:

$$\frac{V^-}{\Delta P^-} = -\frac{1}{\Phi_T}$$

where V and ΔP stand for the transaction volume and price change, respectively. Because $\Phi_T > \Phi_U$, Epps now claims that "... the ratio of volume to price change on upticks exceeds the absolute value of the ratio of volume to price change on downticks" (p. 592). One should note, however, that a necessary condition for the derivation of Epps' equation (19) and (20) is *that both Ψ_T and Ψ_U are parameters and are not functions of P_k*. Tracing back the definition of the Ψ's to equation (14) one might be tempted to agree that, because the summation in equation (14) is for all $i \neq k$, Ψ_T and Ψ_U are not functions of P_k. This is however a fallacy. As can be seen from equation (13) P_k is a function of all P_i $(i \neq k)$. Because this equation holds for all i, it follows that all P_i $(i \neq k)$ are functions of P_k and therefore α_k (equation (14)) is a function of P_k. To express the impact of the price change of security k on α_k we write:

$$\frac{d\alpha_k}{dP_k} = \sum_{i \neq k} \frac{\partial \alpha_k}{\partial P_i}\frac{\partial P_i}{\partial P_k} = \sum_{i \neq k}\left(-\frac{S_k}{p_k}\frac{r_{ik}}{r_{kk}S_i}\right)$$
$$\left[\left(-\frac{S_i}{r_{ii}}\right)\left(\frac{r_{ki}}{S_k} + \sum_{i=1}^{n}\frac{r_{ii}(p_i - P_i)}{S_i}\right)\right]$$

We have no reason to suspect that this expression is equal to zero, neither can it be claimed to be a second-order effect. Hence, any change in expectation of one group will cause a shift in the values of the intercepts from Ψ_U to Ψ_U^* and from Ψ_T to Ψ_T^*.

After recognizing the dependency of the demand intercepts on the price P_i we can write the new equilibrium price after the arrival of a bullish piece of information as

$$P_1 = \Psi_U^*(P_i) - \Phi_U(a_{U0} - V^+)$$

and thus

$$\Delta P^+ = \Psi_U - \Psi_U^* + \Phi_U V^+$$

Likewise,

$$\Delta P^- = \Psi_T - \Psi_T^* - \Phi_T V^-$$

From these last two equations we see that no unequivocal relationship between price change and transaction volume can be specified.

REFERENCES

T. W. **Epps**, "Security Price Changes and Transaction Volumes: Theory and Evidence," *Amer. Econ Rev.*, Sept. 1975, *65*, 586-97.

P. A. **Samuelson**, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments," *Rev. Econ. Stud.*, Oct. 1970, *37*, 537-42.

J. **Tobin**, "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, *26*, 65-86.

Security Price Changes and Transaction Volumes: Reply

By THOMAS W. EPPS*

My 1975 paper presented a model of security prices which implies that the number of shares traded in a transaction, expressed relative to the absolute value of the *change* in price from the previous transaction, is greater for transactions in which price increases than for those in which price falls. Distribution-free tests with data for one month's transactions in each of a sample of corporate bonds supported the hypothesis. The paper by Mark Hanna extends my empirical work and shows that the predicted relation between ratios of volume to price change appears to hold in periods of predominantly falling prices as well as in bull markets.

The comment by Meir Schneller contains two major criticisms of the model in my 1975 paper: 1) that the assumptions are inconsistent with the mean-variance approach; and 2) that, even ignoring this difficulty, the results do not follow from the assumptions.

By the first criticism it is meant that the utility function I adopted is not consistent with the von Neumann-Morgenstern axioms for behavior in the presence of risk; that is, under the assumptions in my paper there is no utility of wealth function (the existence of which is implied by the axioms) whose expected value is a function $U(\mu, \sigma^2)$ of mean and variance alone for which $\beta = -2(\partial U/\partial \sigma^2)/(\partial U/\partial \mu)$ is constant. Actually, the formal assumptions in the paper do admit such a function, and one which is in fact widely used in applied work. Indeed, the function yields exactly the asset demand functions which were derived in the paper (equation (7)).[1]

Consider the negative-exponential utility function, $u(w) = -\exp(-\beta w)$, where w is final-period wealth, equal to $a'X + a_{n+1}$; a and X being n-element vectors of quantities and end of period values (EPVs) of the risky assets, and a_{n+1} being the quantity of the riskless asset. Assumption 6 in the paper requires that the coefficients of variation of the X_i remain constant in the face of news which alters their subjective distributions, which can be satisfied if the X_i are normally distributed with mean p_i and variance proportional to p_i^2. Letting S represent the covariance matrix $E(X - p)(X - p)'$, we have that $w \sim N(\mu, \sigma^2)$, where $\mu = a'p + a_{n+1}$ and $\sigma^2 = a'Sa$. It follows that expected utility is given by

$$(1') \quad E[u(w)] = -\exp(-\beta\mu + \beta^2\sigma^2/2)$$

Writing $U(\mu, \sigma^2)$ for the right-hand side of (1'), it is seen that $-2(\partial U/\partial \sigma^2)/(\partial U/\partial \mu) = \beta$; and maximizing $U(\mu, \sigma^2)$ subject to the wealth constraint, $W - a'P - a_{n+1} = 0$ (where P is the vector of prices of the risky assets), yields, on solving the first-order equations for a, equation (7).

Clearly, this result requires that X_i be normal. In the paper the *lognormal* distribution was cited as an example of a distribution with constant coefficient of variation, although lognormality was not formally assumed. If the X_i are viewed as lognormal, then Schneller is formally correct in stating that the function $U(\mu, \sigma^2)$ with constant β is inconsistent with the expected-utility hypothesis.[2] However, even in the log-

*Associate professor, University of Virginia.

[1] Equations used in this reply are numbered (1'), (2'), etc. Equation numbers without primes denote equations in the original paper.

[2] In fact, Paul Samuelson's well-known approximation theorem cannot be regarded as a defense of my formulation (see Schneller's fn. 3). Samuelson's results and their generalizations by James Ohlson support the use of a quadratic $u(w)$ in cases where the holding period is very short, but my assumption of constant β rules out quadratic utility altogether.

normal case it may be reasonable for traders to consider as an approximation that w, the *EPV* of a well-diversified portfolio, is normal. (See the study by Lawrence Fisher and James Lorie and the conclusions drawn from their results by Jan Mossin.)

Schneller's second criticism, while rather technical, is quite important and is valid (except in one case, to be discussed), although his arguments do not correctly support it. He argues that the main result of the paper—that the ratio of volume to price change on upticks ($V^+/\Delta P^+$) exceeds that on downticks ($V^-/|\Delta P^-|$)—fails to hold because there are feedback effects of price changes in one asset on prices of the other assets. Specifically, he argues that an increase in the price of asset k, P_k, due to a shift in the bulls' demand function would lead to adjustments in all other prices, which would in turn shift the bears' demand functions for asset k. Such shifts would change in an unpredictable way the expression (28) for $V^+/\Delta P^+$; and a similar problem would affect equation (34).

To examine the argument in greater detail, consider an individual's demand function for asset k, expressed in Marshallian form:

$$(16) \qquad P_k = \alpha_k p_k - \gamma_k a_k$$

where

$$(14) \qquad \alpha_k \equiv 1 + \frac{S_k}{p_k} \sum_{i \neq k} \frac{r_{ik}(p_i - P_i)}{r_{kk} S_i}$$

Suppose the demand function (16) belongs to any one of the bears. The question at issue is whether a change in P_k caused by a change in the bulls' expectations alone (i.e., in their p_k's but not in the p_k's of any of the bears) leads to a change in the factor α_k, thus shifting the bears' demand functions, including (16). Schneller argues (a) that the increase in p_k resulting from the actions of bulls will directly affect all the α_i, $i \neq k$, of the bears, since they depend on P_k by equations analogous to (14); (b) that changes in the bears' α_i's will influence the P_i, $i \neq k$, by equations analogous to (16); and (c) that these changes in the P_i will in turn alter α_k for all bears.

Arguments (a) and (c) are correct, but (b) and Schneller's expression for $d\alpha_k/dP_k$ are not, since equation (16) and analogous equations for the other assets are *individual* demand functions, which determine a_1, $a_2, \ldots, a_n$ for one person, rather than P_1, $P_2, \ldots, P_n$. To determine market prices, we return to equation (7), written for the jth of the J traders:

$$(2') \qquad a_j = (\beta_j S_j)^{-1}(p_j - P)$$

where a_j is the vector of the jth trader's holdings of risky assets, and P is the vector of their prices.[3] Summing for all j and equating to the vector A of total asset supplies leads to the solution for market prices:

$$(3') \qquad P = \sum_j \theta_j p_j - \phi A$$

where $\phi = [\Sigma_j(\beta_j S_j)^{-1}]^{-1}$
and $\theta_j = \phi(\beta_j S_j)^{-1}$ are $n \times n$ matrices.

The critical issue in determining whether a change in P_k affects α_{jk} (the α_k of the jth trader), $j = 1, 2, \ldots, J$, is whether changes in some (or all) traders' expected values of the kth asset's *EPV* affect equilibrium prices of the *other* assets; that is, whether changes in any of the p_{jk}, $j = 1, 2, \ldots, J$, affect the P_i, $i \neq k$. From (3') it can be seen that at least some of the P_i do depend on expected *EPV*'s of other assets unless the matrices θ_j are diagonal, and the only simple way for this to occur is for all the S_j to be diagonal, that is, for all assets' *EPV*'s to be uncorrelated. Thus, for the "partial equilibrium" assumption invoked in the paper (p. 590) to hold requires in practice that all assets' *EPV*'s be uncorrelated. Otherwise, it is generally true that shifts in one group's demand function for asset k do lead to changes in prices of other assets and to consequent shifts in the demand function of the other group. While

[3] In (2') each element of the covariance matrix S_j is multiplied by the scalar β_j.

this appears to be a second-order effect,[4] Schneller is correct that the implications about the magnitudes of volume-to-price change ratios do not rigorously follow, except in the absence of perceived correlations among *EPV*'s. Clearly, the model is the worse for requiring this assumption.

[4]For example, when there are just two risky assets and two traders, it is easy to show that the rate of change of one trader's α_k with the other's $p_k (k = 1, 2)$ is of order R_{12}^2, where R_{12} is the correlation between the two assets' *EPV*'s.

REFERENCES

T. W. Epps, "Security Price Changes and Transaction Volumes: Theory and Evidence," *Amer. Econ. Rev.*, Sept. 1975, *65*, 586 97.

L. Fisher and J. H. Lorie, "Some Studies of Variability of Returns on Investments in Common Stocks," *J. Bus., Univ. Chicago*, Apr. 1970, *43*, 99-134.

M. Hanna, "Price Changes and Transaction Volumes: Additional Evidence," *Amer. Econ. Rev.*, Sept. 1978, *68*, 692-95.

Jan Mossin, *Theory of Financial Markets*, Englewood Cliffs 1973.

J. A. Ohlson, "The Asymptotic Validity of Quadratic Utility as the Trading Interval Approaches Zero," in William T. Ziemba and R. G. Vickson, eds., *Stochastic Optimization Models in Finance*, New York 1976

P. A. Samuelson, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances, and Higher Moments," *Rev. Econ Stud.* Oct. 1970, *37*, 537 42.

M. Schneller, "Security Price Changes and Transaction Volumes: Comment," *Amer. Econ. Rev*, Sept. 1978, *68*, 696-97.

The Long-Run Analysis of the Labor-Managed Firm: Comment

By K. V. Berman and M. D. Berman*

In a recent article in this *Review*, Eirik G. Furubotn challenged the traditional theory of labor management as developed by Benjamin Ward (1958, 1967), Evsey Domar, and Jaroslav Vanek (1969, 1970). He concluded on the basis of his own "realistic" model that: "Whatever its contribution to industrial democracy, it [the labor-managed firm] is not an inherently efficient economic organization" (p. 122). Furubotn was correct to criticize the simplistic Ward-Domar-Vanek maximand of money income per worker, which ignores nonpecuniary aspects of utility maximization. This note will show, however, that Furubotn's conclusions for resource allocation do not derive from a revision of the objective function. Rather, they result from a set of unjustified assumptions about the firm's operating environment. With more reasonable assumptions, labor management can yield a Pareto optimal solution for both the short-run and the long-run allocation of resources.

Furubotn, like Vanek, purported to present a generalized partial equilibrium model of a competitive labor-managed economy. The firm was assumed to be a price taker for both inputs and outputs, and was allowed to operate free from external controls. But the competitive labor-managed economy defined by Furubotn's stated assumptions seems a peculiar one indeed. These assumptions include the following:

1) "In a labor-managed socialist system, there is no free capital market..." (p. 119). Investment can be financed only by retained earnings; "...the collective is not permitted to borrow..." (p. 112).

2) "Under labor-management, the value of the firm's capital stock must be preserved in perpetuity..." (p. 116), including additions to the original stock.

3) "Worker-investors ... cannot sell their income rights to others when they leave the organization" (p. 116).

4) However, "...decision makers also have the possibility of owning interest bearing savings deposits..." (p. 115), with fully transferable ownership rights.

5) Labor input in the firm cannot be decreased. It can be varied only by addition of members, and this only up to a maximum number set by the original controlling majority (p. 108). Despite a postulate which states that workers are free from external controls, hours of work are assumed fixed and not within the realm of decentralized managerial choice.

6) Focused myopically on the expansion path of the individual firm, the model lacks any consideration of the impact of the outside world. Despite the purported long-run and dynamic nature of the analysis, Furubotn did not allow for market forces such as entry and exit of firms to move resources toward more highly productive uses (or for any substitute government planning mechanisms to shape the allocation of resources over time).

One wonders how efficient would be resource allocation among profit-maximizing firms in a capitalist economy in which there was no entry or exit; proprietors had no transferable property rights in the firm but could invest in bank savings accounts with full ownership rights; managers were constrained to self-financed investment and prohibited from liquidating unprofitable investment; and managers could vary labor input only by addition of permanent full-time workers up to an arbitrary limit. Such assumptions are surely not essential features of labor management per se, but result from a confusion of Yugoslav institutional de-

*Research associate, Center for Business Development and Research, University of Idaho, and assistant professor, L. B. J. School of Public Affairs, University of Texas-Austin, respectively.

partures from market allocation with a general theory of labor management.

A key premise in the inefficiency finding is Furubotn's assumption that workers would be reluctant to admit additional members. This premise, however, even if accepted, does not in itself lead to long-run resource misallocation, given a capital market (or effective substitute). In an industry with long-run costs characterized by a region of increasing returns, the original majority of a particular enterprise might restrict its size below the economic optimum for this reason. But Furubotn did not assume identical subjective views of optimal membership size by all firms, even of the same initial scale. Some groups of workers might expand their firms to the optimal size because they anticipated pecuniary gains outweighing the disutility of an expanded work force. More importantly, given free entry there would be a strong financial incentive for other groups of workers to form new firms that would enter at a larger scale. Entry of these lower-cost firms would drive the price down toward the minimum long-run average cost of production. The reluctant firms would be forced either to revise their assessment of the balance of pecuniary and nonpecuniary utilities, or go out of business.[1]

Although Furubotn did not specifically mention short-run aspects of resource allocation, he appeared to concur with Vanek's conclusion of inefficient labor use in the short run. If the optimal capital-labor ratio were smaller than the initial position in the firm, worker reluctance to expand membership would be likely to restrict labor input increase to a suboptimal level (p. 122). This result rests on the assumption that the labor input, even in the short run, can be increased only by adding new members.[2]

If, as is argued here, the set of assumptions used by Furubotn is unjustified, what assumptions would be more appropriate for a general model of decision making under labor management?

1) The perfectly competitive market environment of the Vanek model is appropriate for assessing the efficiency properties of an idealized system not tied to the distortions of a particular national economy. Perfect competition specifically includes the opportunity for free entry and exit of firms, and capital market aspects of the perfectly competitive model are not affected by the circumstance of worker management. Long-term loans are available at a market rate of interest. Ownership rights to capital, which may be held by workers individually or by the state (or a combination of the two), include the right of liquidation and the right to a market rental rate and the maintenance of assets when capital is leased out. Where capital is owned by the state and leased out to workers to manage, worker groups have the right to contract with the state for construction and lease of desired plant and equipment at cost, to renegotiate leases at frequent intervals, and to refuse to re-lease inefficient capital assets.[3] An individual worker contributing capital to a firm from prior individual savings or underpaid labor must be compensated by a corresponding individually transferable interest in the firm

[1]Revising the pecuniary-nonpecuniary assessment might take place through a change in the firm's managing worker group. Exit of firms implies here that the inefficient capital stock is to be liquidated by the owners, either worker owners or the state. See fn. 3.

[2]Lack of work-hours variation as a decision option is also the reason for short-run inefficiency in the Vanek model of labor management. Unlike other writers, however, Vanek did not assume that working hours are necessarily unchanging. But worktime in his model (as a component of "effort") varies only as a passive response to the reward for work, not as a conscious managerial choice for adjusting inputs and outputs to changing market parameters (see 1970, ch 12, especially pp 248-52). With a product price rise, for example, Vanek's decision mechanism would cause the firm to reduce labor input (membership), overwhelming any tendency of workers to work longer hours.

[3]If government-owned capital leased to workers is inefficient, producing substandard worker incomes, no worker group will be willing to use the particular stock of fixed assets at a positive rental charge, and the government owner would then have to liquidate the inefficient assets. Workers themselves have the right to liquidate capital for which they have paid full value.

or in his accumulated deferred income, as in the case of individual investment in other assets.

2) The ability of the firm to control and vary hours of work to adjust the labor input should be considered an integral part of worker-managers' decision-making authority.

3) On the other hand, it appears reasonable to assume that the difficulties in admitting or expelling members in a labor-managed firm preclude membership change as a method of *short-run* adjustment of the labor input.[4] For the short run we can assume membership to be fixed, thus defining a short-run decision horizon for the worker-managers.

4) It is reasonable to assume with Furubotn that worker-managers will consider nonmonetary as well as pecuniary utility in decision making. Since the unit of labor input is now a working hour, we postulate that for the individual, utility will in general be inversely related to labor input. We assume here that hours of all workers change proportionately (although they need not be identical). The proportionality requirement is made here for mathematical convenience only, but may be considered to be based on technological requirements of job interdependence, or on egalitarian preferences for equal effort and return.[5]

Differences in workers' income-leisure preferences could pose decision-making problems within a firm, as Robinson pointed out. Furubotn assumed that the firm is controlled by a homogeneous bloc of members from the original decision-making majority. This concept may be appropriate for a very small firm, but appears less so for the larger and more complex enterprises that would be more typical of a substantially worker-managed economy. In the discussion here, worker homogeneity is not assumed, but the choice of hours and other decision problems within the firm are assumed to be solvable through voting procedures.

In a planning period longer than the short run as defined, workers may leave the firm voluntarily for employment opportunities elsewhere that yield greater utility, but they may not be expelled from the firm involuntarily. Subject to this restriction, firms will attempt to adjust membership (as well as capital) to increase utility produced per member, by recruiting new members or failing to replace attrition. While reluctance to expand membership is possible, and an optimum cooperative size may exist, it seems unlikely that workers would in general sacrifice increased income because they are "suspicious of change" to the extent assumed by Furubotn.

The firm's workers decide on inputs of labor time and other resources in order to maximize utility for the typical (ith) worker:

$$(1)\quad \max V_i = u_i(y_i, h_i, n) - \lambda\left[y_i - \frac{\sum pf(H, m, k) - \sum cm - \sum rk}{n}\right]$$

where n is the number of full-time equivalent worker-members; utility u_i is related positively to income y_i, and negatively to hours of work h_i; total hours of all workers H is nh_i; m and k are vectors of current nonlabor inputs and capital inputs, respectively; p, c, and r are exogenously given price vectors of outputs, current nonlabor inputs and capital, respectively; and the Lagrange multiplier λ is the marginal utility of income.

In the short run, the first-order conditions give the maximum of V_i with respect to the current variables m and H (H is varied by changes in h_i, holding n constant):

[4]The ability of enterprises to expel members for temporary financial gain of the remainder has been challenged on practical and philosophical grounds by many writers, including Vanek (1970, pp. 57, 156–58, 385) although his model allows unrestricted firing. See Joan Robinson; James E. Meade; Furubotn and Svetozar Pejovich; Egon Neuberger and Estelle James. With expulsion prohibited or difficult, admitting new members is a long-term commitment for the firm, as well as for the prospective member, that is not likely to be undertaken as an adjustment to short-run conditions.

[5]Theoretical considerations relating to individual choice of working hours in a cooperative are discussed by M. D. Berman.

(2) $V_{h_t} = 0 = u_h/n + \lambda\sum pf_H/n$

or $\sum pf_H = -u_h/u_y$

(3) $V_m = 0 = \lambda(pf_m - c)/n$

or $pf_m = c$ (for each m, c)

These two equations describe the short-run Pareto optimum at which the values of the marginal products of current nonlabor inputs equal their prices, and where the value produced by the marginal labor input equals its opportunity cost. The firm responds to changes in short-run market parameters in the direction of stability, as opposed to the traditional Vanek model. A rise in the price p, for example, increases the remuneration for work and (on the usual assumption of substitution effects outweighing income effects) causes the workers to increase the labor input by increasing the hours worked per member h_t.

In the long run, V_t is maximized with respect to the long-run variables n and k in addition to the two short-run variables h_t and m. Total hours now depend upon n as well as upon h_t. In addition to equations (2) and (3), which still hold in the long run, the first-order conditions for a maximum of V_t are (with the subscript, t, omitted for clarity)

(4) $V_n = 0 = u_n + (\lambda/n^2)[n\sum pf_H h - ny]$

or $-u_h/u_y = (\sum pf_H h - y)/n$

(5) $V_k = 0 = \lambda(pf_k - r)/n$

or $pf_k = r$ (for each r, k)

Firms recruiting or failing to replace members will adjust membership toward the optimum number of workers. If u_n, the utility to the typical worker of an additional worker-member, is zero, the optimum number n in equation (4) is the Vanek equilibrium at equality of the marginal and average value products of labor. Mobility of individual workers toward firms offering higher income levels, as well as entry and exit of firms, will tend to equalize labor incomes among different firms and industries. These market forces will move marginal value products of labor toward Pareto optimal equality throughout the economy. Equation (5) shows the equality of the value of the marginal product of capital with its marginal cost r, which may be either the interest rate (for owned capital) or a rental rate (for leased capital). In the absence of the capital market distortions of the Furubotn model, there is no reason to conclude that worker-managers will use an inefficient scale or capital-labor ratio.[6]

Economists do not ordinarily assess efficiency of an economic system by specifying the institutional distortions of a particular national economy. With appropriate assumptions which are freed from the dominance of Yugoslav institutional aberrations, labor management does not yield so pessimistic a prediction for economic efficiency as Furubotn (and Vanek as well) would have us believe.

[6] Efficiency in the use of capital under labor management was also the conclusion of Meade

REFERENCES

M. D. Berman, "Short-Run Efficiency in the Labor-Managed Firm," *J. Comparative Econ.*, Sept. 1977, *1*, 309 14.

E. Domar, "The Soviet Collective Farm as a Producer Cooperative," *Amer. Econ. Rev.*, Sept. 1966, *56*, 734-57.

E. G. Furubotn, "The Long-Run Analysis of the Labor-Managed Firm: An Alternative Interpretation," *Amer. Econ. Rev.*, Mar. 1976, *66*, 104-23.

——— **and S. Pejovich**, "Property Rights, Economic Decentralization, and the Evolution of the Yugoslav Firm, 1965–1972," *J. Law Econ.*, Oct. 1973, *16*, 275–302.

J. E. Meade, "The Theory of Labour-Managed Firms and of Profit Sharing," *Econ. J.*, Mar. 1972, suppl., *82*, 402–28.

E. Neuberger and E. James, "The Yugoslav Self-Managed Enterprise: A Systematic Approach," in Morris Bornstein, ed., *Plan and Market*, New Haven 1973.

Svetozar Pejovich, *The Market-Planned Economy of Yugoslavia*, Minneapolis 1973.

J. Robinson, "The Soviet Collective Farm as a Producer Cooperative: Comment," *Amer. Econ. Rev.*, Mar. 1967, *57*, 222–23.

Jaroslav Vanek, "Decentralization Under Workers' Management: A Theoretical Appraisal," *Amer. Econ. Rev.*, Dec. 1969, *59*, 1006–14.

———, *The General Theory of Labor-Managed Market Economies*, Ithaca 1970.

Benjamin Ward, "The Firm in Illyria: Market Syndicalism," *Amer. Econ. Rev.*, Sept. 1958, *48*, 566–89.

———, *The Socialist Economy*, New York 1967.

The Long-Run Analysis of the Labor-Managed Firm: Reply

By EIRIK G. FURUBOTN*

In their criticism of my model of the socialist labor-managed firm (1976a), K. V. Berman and M. D. Berman argue that the assumptions underlying the construct are so extreme and specialized as to make its conclusions on allocative efficiency nugatory. The judgement is, in effect, that while my interpretation of labor management may have some relevance for the Yugoslav experience, the theory of the firm presented has no general applicability and can yield no insight into the probable efficiency levels of labor-managed systems based on institutional arrangements different from those assumed for the Yugoslav case. At issue, then, is an important and fundamental question. Is it true, as the authors assert, that if only an "appropriate" institutional structure is posited, a labor-managed system can be shown to be efficient; or does industrial democracy lead inevitably to difficulties of the type exposed in the Yugoslav model?

I. Property Rights and Incentives

The theoretical position taken by Berman and Berman is a popular one and has support from such major writers as James Meade, Jaroslav Vanek, and Jacques Drèze. Nevertheless, I would argue that labor management is, by its essential nature, a system that cannot achieve a Pareto optimal equilibrium. Regardless of the particular institutional environment being considered, if labor management is to exist, it requires the prevailing property rights structure to exhibit certain special features. And this is where the trouble lies. The type of rights structure that permits labor management to function will of necessity generate a perverse pattern of economic incentives—a pattern that leads the system away from Pareto optimality and toward a second best solution.

There are, in short, basic problems common to all labor-managed economies. Thus, despite its limitations, the so-called Yugoslav model is able to point up some *general* principles. Difficulties such as those that arise because workers have finite planning horizons are well exemplified by the Yugoslav case, but they are by no means unique to the Yugoslav institutional structure. We find that labor management always exerts a distinctive effect on the terms under which benefits can be appropriated by decision makers. What has to be recognized is that labor management is at base an artificial construction; its rules are set up with the objective of giving greater economic power to rank and file workers while simultaneously preserving the traditional initiatives, flexibility and efficiency of a decentralized capitalistic system.[1] This value premise, of course, influences structure; thus, an absolutely crucial requirement is that decision-making authority in the firm be restricted to ***current employees***. Institutionally, worker dominance is protected by legal arrangements that limit the holding of claims on the firm's net cash flow to those actually working for the firm. Such claims have no market, however, because to hold a claim one must be employed by the firm and positions in the organization are not legally for sale. The firm as an independent entity has a very modest role; it does not own any durable productive resources and serves merely as a vehicle for worker-initiated transactions. Characteristically, the firm is expected to *rent* capital goods

*Professor of economics, Texas A&M University

[1] Meade offers a concise analytical description of an idealized labor-managed system Implicit in the discussion is the idea that firms other than those that are labor managed are excluded by law

from private individuals who have full ownership rights in them, or borrow capital funds from private creditors.[2]

The preceding sketch of labor management as a "pure-rental" system[3] is consistent with the interpretation offered by Berman and Berman. However, they go on to develop a startling proposition and argue that a competitive market environment of the special type described produces the *same* allocative results as the standard capitalist model.[4] Representative of their thinking is the following telling statement: "... capital market aspects of the perfectly competitive model are not affected by the circumstances of worker management" (p. 702). But this and similar assertions are simply not correct. What the line of analysis ignores completely is the difference in the structure of incentives under labor management. Indeed, allocative difficulties arise precisely because incentives are distorted by the institutional arrangements of the new system.

II. The Planning Horizon and Economic Behavior

According to the rules assumed, each worker in the labor-managed firm has property rights in a particular time stream of income—viz., the stream of net earnings that comes to the firm during the period the worker is employed by the organization. Once the individual leaves the firm, however, he has no claim on the firm's net revenues and no responsibility for the firm's debts or other obligations. This situation is of interest because it has importan implications for behavior. Assuming for th moment that all of the firm's employee have the same planning horizon and expec to remain with the firm for T periods, th forces operating are clear. Incentives ar created for the workers to try to shift cost incurred in the near term to periods after T and, at the same time, to make firm revenues through T as large as possible.

When dealing with leased capital goods workers can be expected to show little interest in protecting and preserving th equipment. Rather, they will tend to us rented assets unsparingly while minimizin or avoiding completely necessary maintenance services. In this way, current employees of the firm can gain greater reward to the horizon and allow future workers t face the negative consequences of capita neglect. Similarly, when borrowing take place, efforts will be made to shift the repayment burden to future groups of workers. In principle, close monitoring of th firm's operations and rigorous enforcement of contracts by capital owners (an lenders) could prevent self-seeking employees from deliberately manipulating th timing of cash inflows and outlays. Bu such procedures are difficult, and in mos cases prohibitively costly.[5] Moreover, mer enforcement of contract would not be abl to stop current decision makers in th labor-managed firm from showing *bias i the choice of production activities.* That is where rates of return are equal, worker will prefer to undertake those activities tha pay off relatively more quickly, or eve those that, while not particularly remunerative over the long term, yield large near-term net revenues.

Rental contracts, then, do not banish th inefficiency problem of labor management.

[2]In the socialist variant of labor management, government agencies are supposed to lease capital goods (and make loans) to firms at rentals that clear the market. This theoretical extension implies that government agencies behave in the same way as private owners of capital. See the author and Svetozar Pejovich, pp. 1137–39.

[3]Drawing on the property rights literature and their own original research, Michael C. Jensen and William H. Meckling (1977) have provided a far-reaching critique of the labor-managed or pure-rental system. Many of their points are incorporated in this paper.

[4]Drèze, p. 1127, also seems to ignore the relation between the property rights structure and behavior.

[5]The fact that, in systems where free choice exists rental of durable production equipment is not alway preferred to ownership suggests that agency costs (se Jensen and Meckling, 1976), can be substantial and that labor management is not an optimal system.

[6]It can be argued that some of the difficulties al luded to above can be avoided if rental contract are of very short (one-period) duration and the cost of renegotiating contracts or repossessing assets ar

Because workers do not have *perpetual* claims on the firm's net cash flows, their concern will be with the appropriation of benefits over the planning interval T. But the special behavior this policy occasions must prevent the attainment of Pareto optimality; indeed, even the moderation of the inefficiencies here would involve large transactions costs avoidable under capitalism.

III. The Entry of New Firms

In the universe envisioned by Berman and Berman, the entry of new firms can be relied upon to drive industry price down to the minimum long-run average cost of production. Thus: "The reluctant [high-cost] firms would be forced either to revise their assessment of the balance of pecuniary and nonpecuniary utilities, or go out of business" (p. 702). This conclusion, however, is not defensible. To achieve a Pareto optimal solution, it is not necessary for all firms in an industry to provide the same working environment or to show the same level of technical efficiency in the production of marketable output. Within limits, any firm can sacrifice wage income (output) for nonpecuniary amenities without endangering its chances for survival.[7] In the case of the labor-managed firm, though, inefficiency will arise unless the members of the collective have identical multiperiod preference functions. This condition holds because, *inter alia*, the politically dominant majority workers can force an undesired wage-environment solution on minority workers. Ideally, labor mobility would insure that each firm had a homogeneous constituency but high search costs under labor management effectively rule out such an equilibrium.[8]

It should also be understood that, in a labor-managed system, the incentives for the formation of a new enterprise are not as powerful as those existing in a normal capitalistic environment. In the latter, an individual seeking entry and organizing a new venture is legally free to capture the present value of the firm's future stream of profits (projected over an infinite horizon). But under labor management the rewards to entrepreneurship are sharply reduced. The worker-entrepreneur does not have unique claim to the residuals generated by the firm; he must *share* the net cash flow each period with the other members of the collective. Moreover, his limited entitlement to income ends when he leaves the firm (for example, at T). And even if the worker-entrepreneur were given some claim to future revenues on his departure, he would have no vote and no effective way of controlling the policies of the firm to protect his "investment." This institutional arrangement implies that there will always be a difference between the perceived *private* benefits of entrepreneurial activity and the true *social* benefits. The result to be expected is of course reduced formation of new firms and inefficient factor allocation.

IV. The Firm's Decision-Making Process

Perhaps the most elusive and difficult problem in the theory of labor management is to establish a persuasive explanation of how workers determine the policies of the firm. We know that all basic decisions are to be made within the organization through some sort of democratic political process. But to say this is not very helpful because there is no widely accepted theory of how such processes operate. In the Yugoslav model, it was assumed that a majority vot-

very low. But even this approach fails because not all of the inputs required by the firm can be rented. As Jensen and Meckling (1977) have explained, intangible organizational factors must be contributed by the workers.

[7]A basic proposition in welfare theory is that a Pareto optimum has not been achieved if a worker can improve his utility level by moving to a lower paid position, see E. J. Mishan, p. 166.

[8]In a labor-managed system, it is quite costly for a worker to make an accurate comparison of the net rewards offered by different jobs (see the author 1976b, pp. 221–24). Further, even with high mobility, the work force at any firm is likely to be heterogeneous because of irreducible age and class differences.

ing procedure permitted a homogeneous subgroup of workers to secure power and, then, run the firm in its own interest. Admittedly, this simple majority rule hypothesis leaves much to be desired. The way in which shifting coalitions of workers may change the firm's policies is not explained, nor is much consideration given to the role of the manager and the means workers have for monitoring and controlling his actions.[9] Nevertheless, the majority rule model does focus attention on the political aspect of labor management, and does highlight the fact that, in any realistic situation where the firm's workers have diverse preferences and planning horizons, a Pareto optimal solution is just not an effective possibility.

By contrast, Berman and Berman appear to ignore the problems of welfare aggregation when they say: "In the discussion here, worker homogeneity is not assumed, but the choice of hours and other decision problems within the firm are assumed to be solvable through voting procedures" (p. 703). There is no doubt that some solution will always emerge; the real question, however, concerns the nature of the solution to be anticipated. Thus, if the employees of a (large) labor-managed firm are generally apathetic and ignorant of policy matters, an energetic and well-organized minority within the collective may dominate the political process. Certainly, there are gains to be realized by those who can control the firm and use its resources for their own purposes. All this points to a simple conclusion; unless a proposed model specifies the political system assumed to be operating within the firm, no clear objective function can be formulated and nothing definite can be said about welfare distribution or allocative efficiency.

[9]Since, under labor management, there is no market for the claims employees have on the firm's net revenue stream, the monitoring of management will be less efficiently performed than in a conventional capitalistic system.

REFERENCES

J. H. Dréze, "Some Theory of Labor Management and Participation," *Econometrica*, Nov. 1976, *44*, 1125–39.

E. G. Furubotn, (1976a) "The Long-Run Analysis of the Labor-Managed Firm: An Alternative Interpretation," *Amer. Econ. Rev.*, Mar. 1976, *66*, 104–23.

———, (1976b) "Worker Alienation and the Structure of the Firm," in Svetozar Pejovich, ed., *Governmental Controls and the Free Market*, College Station 1976.

——— **and S. Pejovich**, "Property Rights and Economic Theory: A Survey of Recent Literature," *J. Econ. Lit.*, Dec. 1972, *10*, 1137–62.

M. C. Jensen and W. H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *J. Finan. Econ.*, Oct. 1976, *3*, 305–60.

——— **and** ———, "On 'The Labor-Managed' Firm and the Codetermination Movement," work. paper, Grad. Sch. Manage., Univ. Rochester, Feb. 1977.

J. E. Meade, "The Theory of Labour-Managed Firms and of Profit Sharing," *Econ. J*, Mar. 1972 suppl., *82*, 402–28.

E. J. Mishan, "A Survey of Welfare Economics 1939–59," *Econ. J.*, June 1960, *70*, 197–256; reprinted in *Surveys of Economic Theory*, Vol. 1, New York 1965.

Jaroslav Vanek, *The General Theory of Labor-Managed Market Economies*, Ithaca 1970.

The Value of Human Life in the Demand for Safety: Comment

By PHILIP J. COOK*

Cost-benefit analyses of public projects designed to reduce mortality from disease or various environmental hazards have virtually without exception used some variant of a labor earnings measure to place a dollar value on lives. This practice continues despite criticism from E. J. Mishan and others,[1] apparently because the necessary data on earnings are readily available whereas data on the theoretically appropriate "willingness-to-pay" measure are incomplete, of recent vintage, and of uncertain quality.[2] This lack of appropriate data does not justify the use of the earnings measure if indeed the latter has no theoretical justification: as Mishan points out in this context "... there is more to be said for rough estimates of the precise concept than precise estimates of economically irrelevant concepts" (p. 705). In the context of this discussion, Bryan Conley's recent attempt (in this *Review*) to provide a theoretical justification for the use of a labor earnings measure of the value of life has great potential significance. Conley's major theoretical conclusion is that "... the value of life saving is greater than discounted lifetime labor income" (p. 51). If true, then Conley's argument could be used to support a claim that if a project's potential "benefits" (as measured by total earnings of lives which would be saved) exceeds the project's costs, then the project is necessarily worthwhile by the cost-benefit standard.

The main purpose of this note is to provide a relatively transparent derivation of Conley's major theoretical result, which avoids the complexities of his multiperiod model.[3]

Consider, as does Conley, an individual with nonhuman wealth W and expected discounted labor earnings Y who lives in a world of full information, competitive, costless markets; in particular, actuarially fair life insurance and annuity contracts are available. At time zero the individual is faced with an endowed probability $(1 - p)$ of immediate death. The appropriate (for cost-benefit purposes) measure of the monetary value of his life can then be inferred from information concerning his willingness to pay for reductions in this probability.[4]

The individual's utility if he survives at time zero is assumed to be a strictly concave increasing function of the expected present value of his lifetime consumption C. It will be assumed that he has no effective bequest motive (i.e., marginal utility of consumption everywhere exceeds the marginal utility of a bequest). He will then choose to purchase an annuity such that $C = Y + W/p$, and his expected utility is given by

$$(1) \quad E(U) = pU(C) + (1 - p)U^* = pU(Y + W/p) + (1 - p)U^*$$

where U^* is defined as "the utility level associated with death" (Conley, p. 51).

The "value of life" is given by the maximum rate at which he would be willing to sacrifice wealth for reductions in p. Taking the derivative of (1) with respect to p while holding $E(U)$ constant yields

*Assistant professor of policy sciences and economics, Duke University

[1] See, for example, the two monographs by Jan Acton.

[2] Acton (1973) reports a survey study of individuals' willingness to pay for a mobile coronary unit which would reduce the probability of death for heart attack victims. See also Richard Thaler, Thaler and Sherwin Rosen, and Robert S. Smith, who attempt to infer the value of life from estimates of risk premiums paid to workers in risky occupations.

[3] My derivation here is similar to the analysis in my paper with Daniel Graham

[4] I ignore throughout the analysis the possibility that other people have an effective demand for his safety.

$$(2) \quad U(C) - U^* + pU'(C) \cdot \left[\frac{1}{p}\frac{dW}{dp} - \frac{W}{p^2}\right] = 0$$

Simplying and substituting $C - Y = W/p$ yields

$$(3) \quad -\frac{dW}{dp}\bigg|_{E(U)=constant} = Y + \left[\frac{U(C) - U^*}{U'(C)} - C\right]$$

Conley bases his conclusion that Y serves as a lower bound for the value of life on the claim that in "... [what] may be called the general case" (p. 50) the expression in brackets is positive. If one assumes (as does Conley in arguing for this conclusion) that $U^* = 0$, then this claim amounts to an assertion that for most individuals the elasticity of utility with respect to lifetime consumption is less than one: for $U^* = 0$, we have

$$(4) \quad -\frac{dW}{dp}\bigg|_{E(U)=constant} = Y + C\left[\frac{1}{\alpha} - 1\right]$$

where α is the lifetime consumption elasticity of utility.

Under what circumstances will $\alpha < 1$? Given the assumption that $U(C)$ is strictly concave, then it is sufficient that $U(0) = 0$. Conley, however, makes the more reasonable assumption that the utility associated with death is equal to the utility associated with some presumably low but nevertheless positive level of consumption, which he labels C^o (see Figure 2, p. 51). Hence, $U(C^o) = U^* = 0$. Then there is some interval for lifetime consumption, $C^o < C < C^*$, for which $\alpha > 1$; clearly C^* is the solution to[5]

$$(5) \qquad C = \frac{U'(C)}{U(C)}$$

[5]Conley (p. 50), demonstrates that equation (5) may have no finite solution, in which case $\alpha > 1$ for all $C > C^o$.

Is it indeed true that "in the general case" individuals have lifetime consumption levels in excess of C^*? The plausibility of Conley's claim in this regard must be evaluated carefully before cost-benefit analysts can feel comfortable with using the present value of lifetime earnings as a lower bound for the value of life.

REFERENCES

J. P. Acton, "Evaluating Public Programs to Save Lives: The Case of Heart Attacks," Rand Corp., R-950-RC, Jan. 1973.

———, "Measuring the Social Impact of Heart and Circulatory Disease Programs: Preliminary Framework and Estimates," Rand Corp., R-1697-NHLI, 1975.

P. J. Cook and D. A. Graham, "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," *Quart. J. Econ.*, Feb. 1977, *91*, 143-56.

B. C. Conley, "The Value of Human Life in the Demand for Safety," *Amer. Econ. Rev.*, Mar. 1976, *66*, 45-55.

E. J. Mishan, "Evaluation of Life and Limb: A Theoretical Approach," *J. Polit. Econ.*, July/Aug. 1971, *79*, 687-705.

Robert S. Smith, *The Occupational Safety and Health Act*, Washington 1976, Appendix B.

R. H. Thaler, "The Value of Saving a Life: A Market Estimate," unpublished doctoral dissertation, Univ. Rochester 1974.

——— **and S. Rosen**, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1975.

D. Usher, "An Imputation to the Measure of Economic Growth for Changes in Life Expectancy" in Milton Moss, ed., *Measurement of Economic and Social Performance*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 38, New York 1973.

The Value of Human Life in the Demand for Safety: Comment

By M. W. Jones-Lee*

In the March 1976 issue of this *Review*, Bryan C. Conley published a paper in which he claimed to have demonstrated, on the basis of a purely theoretical argument, that "... in the general case ... the value of life saving is greater than discounted lifetime labor income" (p. 51). Conley's elegant paper is undoubtedly a substantial contribution to the growing literature on the value of life. Nonetheless, it will be argued that his analytical framework is rather less general and his conclusions less robust, than might initially appear to be the case. More specifically, this note will seek to show (a) that while Conley has (albeit in a somewhat simplified framework) correctly identified a *sufficient* condition on individual preferences for the value of life to exceed human wealth, his theoretical argument contains nothing whatever which would indicate that such a condition is typically met (so that while not denying the possibility that the asserted relation between the magnitude of human wealth and the value of life saving[1] *may* turn out to be correct in a majority of cases,[2] it is my contention that Conley has provided no evidence one way or the other and that the main claim of his paper to have demonstrated the result theoretically is quite without foundation), and (b) that Conley's model suffers from a number of limitations, at least some of which substantially constrain the generality of the condition identified as sufficient for the value of life saving to exceed human wealth.

I. The Methodological Objection

In the context of a conventional scientific program there are essentially two kinds of evidence which can be adduced in support[3] of an empirical generalization: *either* one may provide direct empirical evidence, *or* it can be shown the proposition in question is logically entailed (i.e., necessarily implied) by a "higher level" hypothesis[4] for which supporting evidence has already been provided. On a cursory reading, one could be forgiven for thinking that Conley has adopted the second approach (i.e., demonstration of entailment by a higher level hypothesis) to providing support for the empirical generalization that "the value of life [saving] is greater than discounted earnings" (p. 45). However, a more careful scrutiny of his argument reveals that, following the correct identification of a sufficient condition for such a result[5] (within the

*University of Newcastle Upon Tyne

[1]Conley employs the (by now relatively conventional) definition of the "value of life" as a marginal rate of substitution of wealth for probability of survival. See for example Dan Usher, Richard Thaler and Sherwin Rosen or the author (1974, 1976). Conley then defines the "value of life *saving*" as the sum of the value of life and the excess of lifetime income over lifetime consumption

[2]Indeed, most of the recent empirical work in this field (including work by this author) tends to confirm Conley's assertion concerning relative magnitudes. See Thaler and Rosen, or the author (1976).

[3]Notice that in contrast to logically necessary statements, empirical generalizations concerning invariant associations cannot be *proved* future instances have not yet been observed—but merely *supported* by being shown not to have been disconfirmed by past observations

[4]For a discussion of the structure of science in terms of various levels of hypothesis, see R B. Braithwaite

[5]Conley's condition is that $\partial U/\partial C < U - U^*/C$ where U is the expected utility of lifetime consumption, C is expected discounted lifetime consumption, and U^* is the utility index associated with (immediate) death. This condition is obtained as follows. In order for the value of life saving (value of life *plus* lifetime income, net of consumption) to exceed lifetime income, Y, we require (value of life) + $(Y - C) > Y$; or value of life $> C$. Conley's condition is then obtained by substituting his expression for the value of life, $(U - U^*)/(\partial U/\partial C)$ It is worth pointing out that Conley's expression for the value of life is merely a

terms of his model), Conley then simply asserts that the fulfillment of this condition will be the norm rather than the exception, thus: "... a condition that we expect generally holds for most values of [lifetime consumption].... This critical value is presumably at a low level of income" (p. 50).

Plainly such assertions offer no scientific support whatsoever for the proposition in question.[6] The "... major theoretical finding of [the] paper that the value of life is greater than discounted earnings" (p. 54), is therefore seen not to be a theoretical finding at all, but merely the identification of a sufficient condition together with the unsupported assertion that this condition will normally be fulfilled.

II. Some Limitations of the Analytical Framework

Furthermore, even if we treat Conley's paper as having the more modest objective of identifying the relevant conditions for the value of life saving to exceed human wealth, it seems reasonable to require that the theoretical framework within which this identification is conducted should have wide generality and should not suffer from excessive[7] simplification and abstraction from our perceptions of "reality" (otherwise the conditions deserve to be accorded no more status than that of mere curiosa). Unfortunately, Conley's model of individual choice is based upon at least two very substantial abstractions. In the first place, following the seminal paper by Menahem Yaari, Conley assumes full information, costless, competitive markets for borrowing, lending and life insurance (strictly, for regular and actuarial notes), and as a consequence of this assumption treats the decision-making agent as having immediate access to the full value of expected discounted lifetime labor income (i.e., human wealth). The latter, together with current nonhuman wealth is then treated as the sole ultimate constraint upon an individual's current expenditures including those devoted to safety improvement. However, it is a fact that a variety of market imperfections and information failures conspire to deny many individuals immediate access to more than a small fraction of their human wealth (unless for the purchase of a very limited range of specific durable assets providing good collateral such as housing). Thus the magnitude of currently realizable wealth is for such individuals smaller than is implicitly assumed in Conley's paper. The impact of this is most easily seen in the context of the simplified version of Gary Fromm's single period model (discussed by Conley in fn. 16) in which the value of human life (the marginal rate of substitution of wealth for the probability of survival) is given by[8]

$$(1) \qquad -\partial W/\partial p = U/p(\partial U/\partial W)$$

where W is initial wealth, U is utility of wealth, and p is the probability of survival.

The effect of a *ceteris paribus* variation in initial wealth upon the value of life can then be obtained by partially differentiating $-\partial W/\partial p$ with respect to W, treating p as a parameter:

$$(2) \qquad \frac{\partial}{\partial W}\left(-\frac{\partial W}{\partial p}\right) = \frac{\left(\frac{\partial U}{\partial W}\right)^2 - U\frac{\partial^2 U}{\partial W^2}}{p\left(\frac{\partial U}{\partial W}\right)^2}$$

Conley sets the utility of not living at zero, so that an individual who prefers life to death will necessarily have $U > 0$. Thus,

version of what is by now well established in the literature. (See, for example, Jacques Drèze, equation (14); Usher, equation (9); and the author, 1974, equation (29).)

[6] More explicitly, given the condition identified in fn. 5, one may conclude that if $U'(C) > 0$, $U''(C) < 0$ and $U(C)$ is bounded above, then there will exist *some* value of C, say $\tilde{C}$, such that $C > \tilde{C} \Rightarrow \partial U/\partial C < (U - U^*)/C$. The basic contention of this note is however that Conley offers no reason for supposing that $C > \tilde{C}$ will be the normal case.

[7] Of course what constitutes "excessive" simplification is ultimately a question of personal judgement.

[8] It can be shown that the following argument also applies under substantially more general conditions. See, for example, the author (1976).

if as is usually assumed, $0 < p < 1$, $\partial U/\partial W > 0$, and $\partial^2 U/\partial W^2 < 0$, then

$$(3) \qquad \frac{\partial}{\partial W}\left(-\frac{\partial W}{\partial p}\right) > 0$$

That is, the value of life is an unambiguously increasing function of initial wealth. There remains, however, the crucial question of precisely what is meant by "initial wealth" in this context. Consider an individual for whom the constraint on currently available wealth discussed above is binding (i.e., a person who would, given access to all of his human wealth, plan to increase his total spending in the current period). In such a case, it is clear that currently realizable wealth is the relevant interpretation of W for a discussion of the amount the individual is willing and able to pay for an increase in the probability of survival.[9] It therefore follows from inequality (3) that such an individual will have a value of life lower than that implied by Conley's analysis.

The second limitation of Conley's model arises from his treatment of bequests. The major part of his formal analysis is based upon the assumption of no bequest motive. While he does give brief informal consideration to the possibility that an individual might wish to leave an estate, he fails to notice a most important implication of the opportunity to purchase life insurance: amongst other things life insurance provides a means of substantially augmenting (bequeathable) wealth conditional upon death at the expense of premia payments which reduce wealth conditional on survival. The effect of this is again most easily seen in the context of the single period model employed above, though, of course, it is necessary when discussing the bequest motive to include an explicit utility of bequeathable wealth which will be denoted by $\theta(\omega)$, where ω is bequeathable wealth. The value of life without life insurance is then given by[10]

$$(4) \quad -\frac{\partial W}{\partial p} = \frac{U(W) - \theta(\omega)}{p\partial U/\partial W + (1-p)\partial\theta/\partial\omega}$$

If, however, actuarially fair life insurance is available with premium, x, and net sum assured, y, related by

$$(5) \qquad x = \frac{1-p}{p}y$$

then, provided that initially $\partial\theta/\partial\omega > \partial U/\partial W$, the individual will proceed to purchase life insurance up to the point at which $\partial U(W-x)/\partial W$ just equals $\partial\theta(\omega+y)/\partial\omega$. It can then be shown[11] that the value of life *with* life insurance (which will be denoted by $-\partial W/\partial p\,|_{x\neq 0}$) is given by

$$(6) \quad -\frac{\partial W}{\partial p}\bigg|_{x\neq 0} = \frac{U(W-x) - \theta(\omega+y)}{\partial U(W-x)/\partial W} + (x+y)$$

Comparing (4) and (6) it is clear that, while no unambiguous general restriction is implied concerning the relative magnitudes of the value of life with life insurance and the value of life without life insurance,[12] it is nonetheless perfectly possible for the former to be somewhat less than the latter. Indeed it seems plausible that such a relationship would exist in the majority of cases given that an individual will presumably display a bequest motive largely because he wishes to provide for surviving dependants. By purchasing life insurance, the individual can mitigate the reduction in dependants' income which would otherwise occur should he die and he is thereby able to reduce (to some extent) the distastefulness of the prospect of his own death. This in turn would tend to lower the value which the individual places upon a marginal reduction in the probability of his own death.

It is clear, therefore, that under perfectly

[9]Strictly speaking $U(W)$ then becomes a utility function for initial *realizable* wealth and is well defined only if *either* total wealth (including all of the individual's human wealth) is treated as parametric *or* those expenditure plans which are constrained by realizable wealth are independent of the nonrealizable portion of total wealth.

[10]See, for example, the author (1976).

[11]See, for example, Philip Cook and Daniel Graham.

[12]For an individual who purchases life insurance, $x + y > 0$ by definition. On the other hand, if U and θ are both increasing and concave than $U(W-x) - \theta(\omega+y) < U(W) - \theta(\omega)$.

plausible circumstances either the existence of restrictions on the magnitude of an individual's currently realizable wealth and/or the availability of life insurance will yield a value of life (and hence of life saving) for that individual which is lower than implied by Conley's analysis. Thus, even if we view Conley's paper as having the essentially taxonomic objective of identifying the conditions under which the value of life saving will (or will not) exceed human wealth the above limitations rather restrict the generality of his conclusions.

III. Some Minor Errors and Omissions

In addition to the objections noted so far, Conley's paper also contains some minor errors and omissions which although potentially confusing and/or misleading do not affect his major conclusions. Briefly, these are:

(a) Conley's conditional probability relations are erroneous. He defines p^{Tt} as the conditional probability of survival "through" (presumably to the end of) period t, the conditioning event being that the individual is alive at the beginning of T. He then asserts that

$$p^{0t} = p^{0T}p^{Tt} \tag{7}$$

and

$$p^{0t}_{iT} = p^{0T}p^{TT}_{iT}p^{Tt} \tag{8}$$

where subscripts denote partial derivatives with respect to activity x_{iT}.

Given his definitions, these relationships should, of course, be

$$p^{0t} = p^{0T-1}p^{Tt}(\text{or } p^{0T}p^{T+1t}) \tag{9}$$

and

$$p^{0t}_{iT} = p^{0T-1}p^{TT}_{iT}p^{T+1t} \tag{10}$$

In addition, Conley later refers to p^{TT}_{iT} as a "changed probability." This terminology is grossly misleading in view of the fact that it is of course a rate of change (strictly, a partial derivative).

(b) Conley's treatment of the question of externalities is again potentially misleading. He appears to suggest that within the type of analysis which he is conducting, externalities may be adequately reflected through the presence of some of individual j's activity levels in individual i's utility function ($i \neq j$). What Conley fails to point out is that under conditions of uncertainty, externalities occur whenever there is interdependence between two or more individuals' *expected* utilities and that this interdependence can occur *either* due to interdependent utilities (such as suggested by Conley) and/or due to interdependent *probabilities*. Thus it is quite possible for individual i to be completely indifferent concerning the extent of any of individual j's activity levels per se (i.e., for the latter not to appear at all as arguments of i's utility function) and yet for substantial externalities to exist because of interdependence between expected utilities induced by interdependence between probabilities of survival.[13] The most extreme version of this type of interdependence occurs in those cases in which safety changes possess the characteristics of pure public goods (for example, road safety schemes).

(c) Conley quotes the empirical results of Thaler and Rosen as confirmation of the major theoretical finding of his paper. What he does not do, however, is to mention other published empirical work which is not so encouraging from his point of view. In particular, Stanley Melinek has obtained values of life (as defined by Conley) as low as £28 thousand—somewhat less than average human wealth for working adults in the United Kingdom. Furthermore, it is notable—especially in view of his claims to originality in the early paragraphs of the paper—that Conley does not refer to other theoretical work in the value of life field, broadly founded in the change-in-probability-of-death (or survival) position, some of which predates his paper by more than a decade.[14]

[13] L. Needleman has recently published a paper in which he attempts to estimate the value of improvements in the safety of other people's (typically close relatives) lives using data on kidney transplants. The latter constitute a good example of the kind of direct externalities under discussion since a transplant typically generates a small increase in the probability of death for the donor and a large decrease in the probability of immediate death for the recipient.

[14] For example, see Drèze and Usher.

IV. Conclusion

It should be stressed that the above comments are intended more as a constructive, if somewhat cautionary, critique rather than as a destructive attack upon the core of Conley's argument. It is, in fact, my view that Conley has made a very worthwhile contribution to the value of life literature, filling the gap between existing single period and continuous-time models,[15] though in all fairness it should be noted that Usher has also developed and discussed a multiperiod model, albeit rather less general than Conley's.

This is not the appropriate place to present a detailed comparison of the pros and cons of each of the alternative approaches to the analysis of the value of life (in particular the single period, multiperiod, and continuous-time models). However, it is perhaps worth noting that the Conley-Usher multiperiod consumption models do seem to be vulnerable to the criticism that they accord undue importance to consumption per se. Now it seems reasonable to assume that most individuals would prefer, *ceteris paribus*, to die later rather than sooner. The Conley-Usher analyses seem to allow for such a preference only in the rather specific (and one suspects limited) sense that they make lifetime utility an increasing function of lifetime consumption, the latter being to some extent, dependent on length of life.[16]

[15] For examples of the single period models, see Drèze, Fromm or the author (1974). A continuous-time model is presented in the author (1976).

[16] For example, Usher employs an additive-separable utility function for lifetime consumption having the form

$$U(C_0, C_1 \dots C_n) = \sum_t \frac{C_t^{\beta}}{(1+r)^t}$$

where β and r are utility function parameters. Clearly by setting $\beta = 1$ the individual can be made indifferent to the length of time for which he survives provided that the discounted present value of lifetime consumption is maintained at a particular value.

REFERENCES

R. B. Braithwaite, *Scientific Explanation*, Cambridge 1953.

B. C. Conley, "The Value of Human Life in the Demand for Safety," *Amer. Econ. Rev.*, Mar. 1976, *66*, 45–55.

P. J. Cook and D. A. Graham, "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," *Quart. J. Econ.*, Feb. 1977, *91*, 143-56.

J. Drèze, "L'Utilité Soçiale d'une Vie Humaine," *Rev. Française Recherche Operationelle*, 1962, *23*, 93-118.

G. Fromm, "Comments [on Schelling]," in S. B. Chase, Jr., ed., *Problems in Public Expenditure Analysis*, Washington 1968, 166–76.

Michael W. Jones-Lee, "The Value of Changes in the Probability of Death or Injury," *J. Polit. Econ.*, July/Aug. 1974, *82*, 835–49.

———, *The Value of Life: An Economic Analysis*, London; Chicago 1976.

S. J. Melinek, "A Method of Evaluating Human Life for Economic Purposes," *Accident Anal. Prevention*, Oct. 1974, *6*, 103 14.

L. Needleman, "Valuing Other People's Lives," *Manchester Sch. Econ. Soc. Stud.*, Dec. 1976, *44*, 309–42.

R. Thaler and S. Rosen, "The Value of Saving a Life: Evidence from the Labor Market" in Nestor Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1975.

D. Usher, "An Imputation to the Measure of Economic Growth for Changes in Life Expectancy" in Milton Moss, ed., *Measurement of Economic and Social Performance*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 38, New York 1973.

M. E. Yaari, "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, *32*, 137–50.

The Value of Human Life in the Demand for Safety: Extension and Reply

By BRYAN C. CONLEY*

Both Philip Cook and Michael Jones-Lee criticize my assumption that the critical value above which the value of human life (VHL) exceeds income is "presumably at a low level of income." In reply, I develop an extension to the original article, which eliminates this qualification, and reconstruct my model in light of their comments.

In line with common usage, I had defined income as "expected discounted lifetime labor income" (p. 47) without any subtraction of necessary expenditures as would be required in determining business net income. However, such expenditures must be so subtracted to properly define a meaningful net income. As a leading accounting text emphasizes: "In computing the net income (available for dividends) for a period all forms of expense incurred in the production of such net income must be provided for" (Eldon Hendriksen, p. 62). Put another way, all of business net income is a surplus and may be taxed or distributed without any loss of output (in the short run).

The public finance literature has similarly recognized that "consumption for subsistence should be deducted from [gross] income, regarding such outlays as a cost of production" (Richard Musgrave, p. 171), and "Adam Smith admonished that taxable income should be defined as 'clear' income, or as income above subsistence" (Musgrave, p. 95).

To define income accurately, one must therefore subtract from gross receipts a level of expenditure which allows continued work and makes the individual indifferent between life and death. Only gross receipts above this level are potentially taxable or available for safety expenditures. If a major share of gross receipts is taxed leaving funds less than subsistence (as measured by amount C^o in Figure 2 of the original paper), the person will starve to death, otherwise perish, or commit suicide. Thus any useful definition of "net" income consistent with its inherent meaning as a measure of surplus must subtract from some measure of gross income the expenditures necessary for subsistence living. The same holds for net consumption.

However, subsistence income is difficult to measure and possibly highly variable among individuals. I expect it is on average at a "low" level of income, as argued below. Given this change, my results are valid under all circumstances. That is, if $C' = C - C^o$ and $Y' = Y - Y^o$, $VHL > C'(=Y')$ in all cases. Graphically, the slope of the tangent to any point on a curve, as shown in Figure 2, will always be less than the ray arising out of C^o to the intercept value on U. We have found that $VHL = \{(U/C')/(\partial U/\partial C')\} \cdot C'$, and the factor in brackets is the inverse of α'. Since it is also the slope of the ray divided by the tangent, from the previous line of reasoning, the quotient is always greater than one, α' is always less than one and the value of human life is always greater than consumption (where consumption and income are defined net of subsistence expenditures).

Cook wishes "to provide a relatively transparent derivation of Conley's major theoretical result, which avoids the complexity of his multiperiod model" (p. 710). A simple model of the main results is contained in my original footnote 16.[1] His last sentence is a reasonable caveat to my origi-

*Graduate School of Business Administration, University of Southern California. I appreciate comments by Dennis Capozza, Hartry Field, and William Schulze, but responsibility for any error is mine.

[1] My simple model avoids his introduction of W/p. This term is eliminated in his equation (3) and only strengthens my results if it is positive. His equation (4) is contained on page 51.

nal model but one which available data supports.[2]

Jones-Lee's comments can best be addressed by redevelopment of the paper. The major conclusion is a theoretical conditional statement: "for a value of expected lifetime consumption [equal to income] above some critical value where $\alpha^1 = 1$, the value of human life exceeds expected lifetime consumption" (p. 50). The validity of this statement is not questioned, but instead, objection is raised to the plausibility of assumptions. All models are validated through the accuracy of their predictions, the positivistic canon. On this account, my results have been confirmed.[3] If all theory had to have its assumptions independently confirmed, we would have to dispense with most of economic theory.

While the motivation for exploring the relationship between *VHL* and income was related to public policy, the purpose of the paper was stated in the introduction to be a theoretical paper describing a model which links *VHL* with wealth and utility function characteristics.

The assumptions are standard for a choice-theoretic model designed to include uncertainty. One can adopt different sets of assumptions, some possibly more realistic, but clear-cut results will not emerge—a desired goal. It is still useful to state the "identifying relevant conditions" from which a definite conditional statement arises.

Of the dozen or so assumptions made, alternative specifications of the model would have led to higher or lower determinations of *VHL*. However, these are different models. Certainly an endowment of material wealth ($W^o > 0$) would increase *VHL*. Several other specifications would also lead to higher values.

The model does not give "undue importance to consumption per se" (Jones-Lee, p. 716) as the individual's ultimate objective is expected utility and components of consumption are only part of the activities over which the individual has control. Length of life is not an independent objective, but derived as part of the maximizing process. People will generally strive for greater longevity as it reallocates expenditures from marginal consumption of relatively low marginal utility at younger ages to inframarginal consumption of high and low marginal utility at an older age.

The model permits an individual to borrow against future earnings, as Menahem Yaari and Nils Hakansson assumed in similar theoretical papers. Such borrowing is feasible given the initial assumptions. Of course, "market imperfections" impede the practicality of extensive borrowings, but one can refer to the use of credit cards, student and other forms of unsecured loans.[4] In practice, the level of safety expenditures is most likely a small fraction of one's income (not initial wealth); therefore, financing safety expenditures is not crucial. The magnitude of expected future earnings will have its main effect on the choice between early consumption and safety expenditures which increase survival probabilities. One must not confuse income with wealth. If an individual starts with zero initial wealth, but plans to save some of his income, he has the funds available for safety expenditures irrespective of his ability to tap human capital.

The amended model which includes a bequest motive contains the assumption of actuarial notes. These notes protect beneficiaries against an early death, and thus are life insurance. If without life insurance the distastefulness of death is higher, then the conclusion is further supported that $VHL > C$.

The most cogent of the comments is that I did not provide any support or reason for assuming that the critical value is "low."

[2] See Richard Thaler and Sherwin Rosen, Robert Smith, Stanley Melinek, and studies cited by Jones-Lee (1976).

[3] To which he has no objections (his fn. 2). His own empirical results indicate a value in excess of £3 million (1976, p. 150). While Melinek states that "£50,000 would be a representative implied value of life" (p. 108), his Table 2 indicates that maximum discounted earnings would be £28,800.

[4] Indeed, Jones-Lee (1976, p. 88) indicates that some human capital is marketable

The original reason for ignoring what is certainly the key controversial issue of the article is that it is, strictly speaking, unsupportable. Witness my Figure 2. It was, however, my supposition that most economists would recognize through indirect evidence that most utility curves, if their shape is at all measurable, exhibit considerably more curvature than Figure 2 offers.

The magnitude of the critical value logically depends upon the level of consumption (income) equivalent in utility to being dead and the curvature of the utility function above C^o. We observe many people living at very low levels of income both in this country and abroad. Some derelicts must be living at 10 percent of this country's median per capita income ($7,000) and many persons overseas live on considerably less. These people choose life over suicide because life is preferred. Almost all of the most destitute still cling to life. While some wealthy persons commit suicide, their behavior exhibits depression for other than monetary reasons. I interpret the above as an indication that C^o is a very low (by American standards) level of income.

The degree of curvature of the utility function above C^o is harder to establish. However, the progressivity of the income tax schedule has often been interpreted as justified on account of the diminishing marginal utility of income. In addition, we observe that relatively poor people are quite willing to accept a considerable load above the actuarially fair payment in their insurance premiums.

By way of illustration, consider a person with the quadratic utility function $U(Y) = -1 + 2.5Y - Y^2$, where income is measured in thousands of dollars per year. Since $U(1/2) = 0$, $500 is subsistence income. Also, the critical value Y^* is $1,000. Let us suppose that the individual has half a chance of annual subsistence income of $500 (for a lifetime) and half a chance of $1,000 annual income (for a lifetime). The actuarial value is $250 above the assured $500 income. In lieu of this, he would accept an assured income as low as $691 per year if he maximizes expected utility. Thus, he gives up $59 of a $250 no load premium, for a load percentage of 23.6 percent.

Since many poor people do buy insurance with comparable loads, we may conclude that the utility function is not nearly linear with income. The critical value is at most a small multiple of subsistence income, and therefore below the actual income of most people, at least in developed countries.

Even if the critical value should not be low, VHL must be close to Y. Since $VHL = (Y - Y^o)/\alpha'$ and $\alpha' < 0$, VHL can be less than Y by a maximum of Y^o, which as previously indicated is not large.

In summary, the comments of Cook and Jones-Lee have afforded me an opportunity to further clarify my model, thereby providing some limited support to the assumption that the critical value is indeed low. Any further thoughts are actually well expressed in chapter 7, "Summary and Postscript," of Jones-Lee's book, to which I recommend the reader's attention.

REFERENCES

B. C. Conley, "The Value of Human Life in the Demand for Safety," *Amer. Econ. Rev.*, Mar. 1976, *66*, 45–55.

P. J. Cook, "The Value of Human Life in the Demand for Safety: Comment," *Amer. Econ. Rev.*, Sept. 1978, *68*, 710–11.

N. H. Hakansson, "Optimal Investment and Consumption Strategies under Risk, an Uncertain Lifetime, and Insurance," *Int. Econ. Rev.*, Oct. 1969, *10*, 443–66.

Eldon S. Hendriksen, *Accounting Theory*, Homewood 1977.

Michael W. Jones-Lee, *The Value of Life: An Economic Analysis*, Chicago 1976.

———, "The Value of Human Life in the Demand for Safety: Comment," *Amer. Econ. Rev.*, Sept. 1978, *68*, 712–16.

S. J. Melinek, "A Method of Evaluating Human Life for Economic Purposes," *Accident Anal. Prevention*, Oct. 1974, *6*, 103–14.

Richard A. Musgrave, *The Theory of Public Finance*, New York 1959.

Robert S. Smith, *The Occupational Safety*

and Health Act, Washington 1976, Appendix B.

R. Thaler and S. Rosen, "The Value of Saving a Life: Evidence from the Labor Market," in Nestor Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 40, New York 1975.

M.E. Yaari, "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, *32*, 137–50.

Inflation in Britain: A Monetarist Perspective: Comment

By George Fane*

David Laidler recently presented a monetarist perspective on British inflation in this *Review*. His analysis incorporated two distinctively "monetarist" propositions: first, the adaptive expectations version of the natural rate hypothesis (*NRH*), according to which there is no permanent tradeoff between inflation and unemployment, since events which cause unemployment to fall below the natural rate of unemployment (*NRU*) also cause the actual inflation rate to exceed the expected inflation rate, and vice versa; and second, the claim (p. 487) that, at least since 1968, variations in the rate of monetary growth have caused short-run variations in real activity and unemployment. Combining these propositions Laidler argued that variations in the rate of monetary growth determine variations in the rate of inflation. His article attempted to use this model to explain the variations in inflation and unemployment in Britain since 1953, and in particular to diagnose the causes of the failure of Keynesian policies after 1967 despite their apparent earlier success. In explaining Britain's recent experience of high inflation and historically high unemployment Laidler did not attribute any role to cost-push forces associated with union militancy or with rises in import prices. This note criticizes Laidler's analysis on four grounds: 1) he underestimates the difficulties of reconciling *NRH* with the facts; 2) contrary to the second of the two distinctively monetarist propositions listed above, variations in the rate of growth of the money supply have not been closely correlated with changes in real activity; 3) his explanation for the failure of Keynesian fiscal policies after 1967, despite their apparent success before 1967, is wrong; and 4) Laidler is wrong to minimize or deny the contribution to Britain's depressing record of high inflation and high unemployment in the 1970's, of increases in union power and militancy and of increased world prices of imported raw materials. These points are dealt with in the next four sections. The conclusions are given in the last section.

I

Contrary to the impression given by Laidler (pp. 487, 490–91, 495–96) *NRH* has not explained Britain's inflationary experience: his Figure 1 shows that there has been a long-run increase in unemployment over the period 1953–75, accompanied by a long-run acceleration of inflation. The introduction of the earnings-related supplement to unemployment benefits in October, 1966 presumably did raise *NRU*. However Laidler's Figure 1 shows that even within the subperiod 1967–75 there has been a general upward trend in unemployment, accompanied by a sharp acceleration in inflation; and in the subperiod 1953–66 there was a slight upward trend in unemployment but no clear trend in the rate of inflation. Given Laidler's assumption that expectations are formed adaptively, changes in the rate of inflation can never be fully anticipated; under these conditions *NRH* predicts that changes in the rate of inflation will be inversely correlated with the level of unemployment. Thus for the whole period, and even within each subperiod, the observed movements in unemployment and inflation conflict with the predictions of *NRH*. Laidler however "explains" the facts by introducing whatever *ad hoc* adjustments to *NRH* are necessary in order to explain them. When the expectations augmented

*Australian National University. I am grateful to Eric Kiernan and the referees for several helpful comments. David Laidler generously pointed out some of the faults in an earlier draft.

Phillips curve appears to shift, he appeals to the way in which the 1967 devaluation, the Vietnam War inflation, and the destruction of the Bretton Woods system impinged upon Britain (p. 487); the unusually slow growth of real wages in the mid-1960's (p. 490); the short-term effect of wage and price controls on wages in 1966-68 (p. 490); lags in the adaptation of expectations to an accelerating and largely imported price inflation (pp. 490-91); an increased dispersion of demand both between industries and regions during the 1972-73 boom (pp. 495-96); demographic factors associated with the low birth rate of the interwar years coupled with the high postwar birth rate (p. 496); the "shake out" of formerly hidden unemployment by firms during the recession of 1970-71 (p. 496); and the increases in unemployment benefits in the mid-1960's (p. 496). Although there have been frequent changes in most of these factors during the postwar period Laidler never considers their possible relevance on the various occasions when *NRH* gives correct predictions. Nor is an extensive list of adjustments wholly redundant if one wishes to reconcile *NRH* with the facts, since it appears to be impossible to explain the necessary shifts in the Phillips curve from the changes in unemployment benefits alone: Laidler refers to the study by Dennis Maki and Z. A. Spindler, but their estimates suggest that the increase in unemployment benefits from the average value for the period 1953-65 to the average for the period 1967-72 raised the unemployment rate by less than 1 percent, whereas the required rise in *NRU* is 2 percent or more; the Phillips curve estimates of Malcolm Gray, Michael Parkin, and Michael Sumner suggest that *NRU* rose from 1.8 percent in the late 1950's and early 1960's to 3.7 percent by 1968; Gray, Parkin, and Sumner also indicate (p. 43) that the estimated *NRU* for the post-1968 period would have been even higher than 3.7 percent if they had not included in their Phillips curve a separate dummy variable for the period 1971-72. Indeed the inclusion of this variable implies an estimated *NRU* of 5.6 percent for 1971-72 (assuming an underlying trend rate of productivity growth in this subperiod equal to that for the full period 1968-74). By assuming that *NRU* is a function of unemployment benefits it should in principle be possible to eliminate dummy variables from the Phillips curve and to test whether changes in these benefits can explain the shifts in the Phillips curve which are needed to reconcile *NRH* with the data. This was attempted by Parkin who included a benefits variable and a 1966 shift dummy, however he found both variables statistically insignificant.

II

Laidler claims (pp. 486-87) that there has been a particularly marked pattern, from 1968 onwards, between monetary expansion and inflation, with a two-year lag. He argues that the transmission mechanism for this process involves the Phillips curve. Laidler appears to believe that the lag between monetary changes and changes in unemployment is about one year,[1] which would imply another lag of about one year between unemployment and price inflation. However, given Laidler's transmission mechanism, one should *not* expect to observe the regular two-year lag between monetary expansion and inflation in the period after 1968 since the Phillips curve "... vanishes between 1967 and 1971 ..." (p. 487). Careful inspection of Laidler's Figure 1 reveals that the regular two-year lag is preserved only because both parts of his transmission mechanism broke down in this period: the disappearance of the Phillips curve was offset by the fact that the unemployment rate increased from late 1970 until early 1972, despite an upturn in the rate of growth of the money supply which began in late 1969 or early 1970 and continued beyond early 1971.[2] To preserve

[1] He attributes the rising unemployment of 1974-75 to the monetary contraction of 1973-74 (p. 499).

[2] Superimposed on the very large increase in the rate of growth of M_3 between 1969 and 1972 there was admittedly a small dip in 1971. If one expected to observe a one-year lag between changes in monetary

Laidler's suggested transmission mechanism it would be necessary to argue that the lag between monetary changes and real changes was much longer in the case of the upturn in the rate of monetary growth, which began in late 1969 or early 1970, than it was in the case of the monetary contraction which began in 1973. In the former case unemployment only began to fall in 1972, whereas in the latter case it began to rise in early 1974. A more plausible explanation for the upturn in unemployment in 1974, given the reduction in real disposable income caused by the deterioration in the terms of trade in 1973–74, is that the demand for imported fuels and raw materials is inelastic and that the demand for domestic goods therefore fell. This contractionary effect more than offset the lagged effects of the expansionary monetary and fiscal policies of 1972 and early 1973.

III

Laidler's explanation for the failure after 1967 of the policy of pursuing high employment with fiscal policy, despite the apparent success of this policy before 1967 is that

> A fixed exchange rate and a low inflation rate in the world economy lay at the root of the apparent success of Keynesian policies in Britain before 1967. These policies led to the devaluation of 1967, which coincided with the beginning of the Vietnam War inflation that ultimately destroyed the Bretton Woods system. The way in which these changes impinged upon Britain accounts for the temporary disappearance of the Phillips curve after 1967. [p. 487]

This explanation is somewhat obscure. What is clear is that a rational expectations model would provide no reasons for supposing that a fixed exchange rate and a low rate of inflation in the world economy are either necessary or sufficient for the maintenance of low rates of domestic unemployment and inflation: in an open economy with a freely floating exchange rate domestic inflation is a weighted average of the actual and expected rates of domestic monetary expansion.[3] If the exchange rate is fixed then the domestic inflation rate is a weighted average of the actual and expected rates of domestic and foreign monetary expansion. In both cases unemployment only differs from *NRU* to the extent that the relevant rates of monetary expansion are incorrectly estimated. In contrast to these results Laidler sets out an adaptive expectations model (pp. 487–88) according to which there is a quasi long-run tradeoff between unemployment and inflation in an open economy with a fixed exchange rate: in his model it is possible for the authorities to keep unemployment below *NRU*, at the expense of having a somewhat higher rate of domestic inflation than the world rate. Eventually such a policy will deplete foreign exchange reserves, but in the meantime unemployment can be kept systematically below *NRU* without causing an acceleration of domestic inflation. These results depend crucially on the mechanistic way in which expectations are formed in Laidler's model, since in a rational expectations model there is no systematic way in which the authorities can keep unemployment below *NRU*, provided that there is no way of inducing people systematically to underestimate the relevant rates of monetary expansion.

Laidler claims that "... over the period 1953–67, too low an unemployment target was set ..." (p. 488). It is much more plausible to argue that the key to the

growth and changes in unemployment one would therefore have expected to observe a large decline in unemployment between 1970 and 1973, interrupted by a small increase in 1972. Seasonally adjusted unemployment actually rose throughout 1970–71, peaked in the first quarter of 1972, and then declined until late 1973.

[3] Changes in fiscal policy are ignored. Such changes would complicate the exposition of the properties of the models discussed in this section without affecting the essence of the conclusion that under rational expectations, and assuming that all factors are supplied inelastically, only unanticipated policy changes affect real activity and unemployment.

lengthy pre-1967 success of fiscal policy was that the unemployment target was usually set about equal to *NRU*. Had the authorities attempted to keep unemployment below *NRU*, Britain could never have maintained a fixed exchange rate for so long, given the low rate of inflation in the world economy.

To explain the bad performance of the British economy after 1967 one must explain not only the acceleration of inflation but also the rise in unemployment. This can be done by postulating an increase in *NRU*; but this increase can not plausibly be attributed to devaluation, the Vietnam War inflation, or the collapse of the Bretton Woods system as implied by the above quotation. Rapid inflation was inevitable in the period after 1967, given the actual rate of monetary growth. However, had *NRU* not increased sharply during this period, the rapid monetary and fiscal expansion would have induced a temporarily low level of unemployment, even by historical standards. Nor were the inflationary policies which Britain has pursued since 1967 a mere aberration; they were a response to the pressures on successive governments to use monetary and fiscal expansion in vain attempts to restore the low levels of unemployment to which the electorate had become accustomed in the period before 1965, despite the subsequent increases in *NRU*. Therefore the key to the initial successes and subsequent failures of postwar demand management policies in Britain does not lie in the exchange rate system, or the rate of world inflation, but in the initial agreement and subsequent divergence between *NRU* and the government's target rate of unemployment.

IV

There is a sense in which Laidler's position is similar to that of the exponents of cost-push inflation whom he criticizes: an exogenous increase in *NRU* will have similar effects to those usually attributed to cost-push inflation. Admittedly some of the proponents of cost-push inflation appear to believe, wrongly, that a permanent process of inflation can occur without sustained monetary expansion. However the cost-push arguments can be rephrased in *NRH* terms: Marcus Miller did this for the case in which import prices rise while the terms of trade deteriorate, using John Hicks' (1974, ch. 3; 1975) model of real wage resistance, and it is obvious that a change in union power or attitudes can alter *NRU*. Both these factors could raise unemployment and the price level, even if the money supply were held constant. In my view these factors are more plausible than many of Laidler's special explanations: Miller's analysis shows the importance of the interaction between import prices and wage indexation in 1974 and it seems hard to deny that the power and militancy of the trade unions have increased in the last decade: the picketing of power stations and the use of the "work to rule" are techniques which have been refined during this period, and successive government attempts to reform the legal framework of industrial relations have been defeated by the unions.

Finally, if expectations are formed rationally rather than adaptively, an anticipated monetary or fiscal stimulus can be immediately offset by higher wages and prices. This hypothesis could conceivably explain the acceleration of inflation despite the absence of abnormally low unemployment, but can not by itself explain the conjunction of rising unemployment and accelerating inflation. This conjunction might be explained, in part, by postulating that unions expected governments to offset the otherwise contractionary effects of deteriorations in the terms of trade by even more expansionary policies than were in fact adopted. Laidler does not refer to the rational expectations hypothesis. He does refer to Hicks and to the many sociological theories of increased union militancy. However all these references are either openly or implicitly critical.

V

The natural rate hypothesis can explain Britain's experience of inflation and unemployment only by postulating several *ad*

hoc adjustments to the basic theory. The result is not a real explanation, since the required adjustments must more than offset the basic prediction of the simple theory that, following a long period of fairly steady inflation and unemployment, a secular increase in unemployment will reduce the rate of inflation. The various sociological and cost-push explanations are also *ad hoc*; indeed they are indistinguishable from some versions of *NRH* in the sense that the main ideas behind the sociological and cost-push explanations can be rephrased in *NRH* terms. The resulting explanations of British experience then amount to *ad hoc* explanations for the presumed increases in *NRU*. The controversy over the explanation of Britain's experience can therefore be reduced to two questions: firstly, should the various sociological and cost-push explanations be rephrased in *NRH* terms? If so, which explanations for shifts in *NRU* are most plausible? I accept the theoretical argument according to which a long-run tradeoff between inflation and unemployment must involve permanent money illusion. This has been denied by James Tobin; however in the present context, the discrepancy between theoretical predictions and facts becomes even worse if one expects to observe a permanent tradeoff between inflation and unemployment. For these reasons I am convinced that the sociological and cost-push theories should be rephrased in *NRH* terms. On the second question, I agree with Laidler that more generous unemployment benefits probably did raise *NRU*; however, it seems implausible to deny or to seek to minimize the importance of either increased union power and militancy or the effects of large increases in the world prices of imported raw materials relative to the world prices of the manufactured goods with which British exports compete. In our present state of knowledge these three reasons for adjusting the predictions of *NRH* offer the least unsatisfactory explanation of Britain's experience of inflation and unemployment. However, confidence in the predictive and genuinely explanatory power of *NRH* could only be restored by an empirical Phillips curve study which explained any apparent shifts in the curve, not by introducing dummy variables, but rather by explicitly measuring and incorporating the factors which are presumed to cause these shifts.

REFERENCES

M. R. Gray, M. Parkin, and M. T. Sumner, "Inflation in the United Kingdom; Causes and Transmission Mechanisms," Soc. Sci. Res. Center, Inflation Workshop paper no. 7518, Univ. Manchester 1975.

John R. Hicks, *The Crisis in Keynesian Economics*, Oxford 1974.

———, "What is Wrong with Monetarism," *Lloyds Bank Rev.*, No. 118, Oct. 1975, 1-13.

D. Laidler, "Inflation in Britain: A Monetarist Perspective," *Amer. Econ. Rev.*, Sept. 1976, *66*, 485-500.

D. R. Maki and Z. A. Spindler, "The Effect of Unemployment Compensation on the Rate of Unemployment in Great Britain," *Oxford Econ. Pap.*, Nov. 1975, *27*, 440-54.

M. H. Miller, "Can a Rise in Import Prices Be Inflationary and Deflationary? Economists and U.K. Inflation, 1973-74," *Amer. Econ. Rev.*, Sept. 1976, *66*, 501-19.

M. Parkin, "Income Policy: Some Further Results on the Determination of the Rate of Change of Money Wages," *Economica*, Nov. 1970, *37*, 386-401.

J. Tobin, "Inflation and Unemployment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 1-18.

Inflation in Britain: A Monetarist Perspective: Reply

By DAVID LAIDLER*

George Fane uses adjectives such as "*ad hoc*," "obscure," "mechanistic" not to mention "wrong" to characterize aspects of my discussion in this *Review* of recent British inflation; so it is safe to assume that he disapproves of it. However it is one thing to express disapproval, and another thing to justify it. Fane does not back up his assertions of disagreement with convincing arguments.

To begin with I quite agree with Fane that the natural unemployment rate hypothesis, combined with an application of the adaptive expectations mechanism to the generation of the expected domestic inflation rate, will not explain recent British experience. I never claimed that it would. It is clear that, if a monetarist explanation of inflation is to be reconciled with the facts of the behavior of the price level and the unemployment rate in the British economy since 1967, two things must be shown to have happened. First it must be shown that the expected inflation rate accelerated after 1967 independently of the past behavior of the domestic price level. Secondly it must be shown that the natural unemployment rate of the British economy increased. My argument proceeded along just such lines. I pointed to several factors as being responsible for these changes. It is refreshing for a monetarist to be criticized for having adopted a multicausal explanation of a series of events and that is what Fane does. In particular he remarks that "although there have been frequent changes in most of these factors during the postwar period, Laidler never considers their possible relevance on the various occasions when the natural rate hypothesis gives correct predictions" (p. 722). Both assertions are inaccurate.

*University of Western Ontario

I attributed the step-up in inflation expectations to the devaluation of 1967, and the simultaneous acceleration in the world inflation rate. The devaluation was a unique event in the post-Korean War period, and its effect on inflation expectations has been thoroughly investigated by other workers (see John Carlson and J. Michael Parkin). Data presented in my paper showed that the acceleration in the world inflation rate that began in the late 1960's was a new phenomenon, completely distinct from anything that had happened before. Moreover the effect of the world inflation rate on British inflation expectations, as well as on those in a number of other countries, was investigated in much more detail for the entire post-Korean War period by Rodney Cross and the author.

Among factors affecting the natural unemployment rate, I referred to changes in unemployment benefits introduced in 1966. Since Fane seems to agree that these did affect the natural unemployment rate, there is no need to comment on this matter further. However I also referred to changes in the demographic structure of the labor force that took place in the late 1960's. J. I. Foster shows that such changes did take place, and were unique in postwar history. According to Jim Taylor's work (for example, 1972) the shake out of formerly hidden unemployment was a phenomenon that became important for the first time at the turn of the decade. Only when it comes to the matter of the role of the dispersion of demand between industries and regions during the 1973 boom is there any basis to Fane's assertion that I have failed to consider the relevance of such a factor at other periods, but even here it is worth noting that the National Institute found the data on this phenomenon striking enough to warrant extensive comment in

the November 1973 issue of their *Review*.

I also dealt with the behavior of real wages during the period 1969-71. I suggested that their rapid growth over these years might be explained in terms of a catch-up effect after unusually slow growth in the middle 1960's, which might partly have been attributable to short-term effects of wage controls. Here, I presented data to show that the unusually slow growth of real wages in the mid-1960's was something which had not been experienced in earlier times. Moreover the study which I cited in support of my claim that wage controls had been effective in 1966-68 (see Taylor) gave no indication that such controls had had a similar influence during earlier periods. In short, on every factor, save one, that Fane mentions, I either presented data myself, or cited a study which presented data on the influence of that factor at other times, data which in each case showed it to be playing a special role in the 1967-72 period.

Fane challenges my suggestion that from 1968 onwards the time path of the inflation rate in Britain can largely be explained in terms of the behavior of the monetary expansion rate. He asserts that "given Laidler's transmission mechanism, one should not expect to observe the regular two-year lag between monetary expansion and inflation in the period after 1968 since the Phillips curve 'vanishes between 1967 and 1971'" (p. 722). But of course what Fane refers to as "Laidler's transmission mechanism" runs from money to *excess demand* to the inflation rate, given the state of inflation expectations. The role of measured unemployment is to act as a proxy for excess demand. Since, as I have already explained, I believe that the relationship between unemployment and excess demand was changing during the period 1967-71, there is no particular reason to believe that we should not observe a conventional money-inflation rate relationship over this period despite the fact that the unemployment-inflation rate relationship was shifting. The behavior of the unemployment rate during the period 1967-72 is not a good indicator of the behavior of excess demand.

Fane is on stronger ground when he notes that the lag between the upturn in monetary expansion in 1970 and that in real activity in 1972 is longer than that manifested in the subsequent downturn. Of course a monetarist should not be too troubled to find that there is a longer lag between a monetary change and an upturn in real activity than between a monetary change and a downturn. That state of affairs conforms qualitatively to the "stylized facts" about the nature of the lag between monetary policy and economic activity. Nevertheless, further reflection has persuaded me that there might be some substance to Fane's concerns about the timing of money-real activity-inflation rate changes during this period. In retrospect, I might have been better advised to distinguish between changes in the behavior of the money supply that were coming through the balance of payments, and changes that were stemming from domestic credit expansion during the first part of this period. It was not until 1972 after all that Britain adopted something resembling a flexible exchange rate. It is clear that the upturn in the behavior of the British monetary expansion rate that began in 1970 was initiated by the balance of payments and only subsequently kept going by domestically generated credit expansion. Thus it may be true that the data which I presented did artificially prolong the lag between monetary and real changes over the years 1970-72. Further work on this episode is obviously called for.

As to the downturn that began in late 1974, I continue to find the attribution of this by Fane, and for that matter by Marcus Miller, to the effect on real income of a deterioration in the terms of trade largely associated with the oil price increase of 1973 extremely implausible. The development of the North Sea oil fields was well underway by 1973 and this implies that the change in oil prices did not have the same effect on British *permanent* income as on current income. Indeed it may well have increased permanent income, which is surely the relevant income variable as far as aggregate

demand determination is concerned. I readily concede that this is not a matter which can be settled by assertion and counterassertion. Not the least of *OPEC*'s sins was to spoil a crucial experiment on the nature of the inflationary mechanism in the British economy by introducing an extraneous factor whose presence makes it difficult, but I hope not impossible, to carry out the empirical historical work which is required to settle the matters at stake here in a fashion that will convince people on both sides of the debate. Before such work can be carried out though, those who like Fane reject the monetary explanation of the 1974 downturn will have to explain what it is about their view of the way that money affects economic activity that leads them to believe that a fall in the monetary expansion rate (M_3) from a level close to 30 percent per annum in early 1973 to less than 15 percent per annum from late 1973 onwards would not have led to a severe recession in the absence of the oil price shock. Until they do, monetarists like myself, who predicted the downturn *before* it occurred and (pace Miller) called for *more* not *less* expansionary monetary policy in early 1974 will remain at a loss to know what evidence to confront them with (compare my evidence in House of Commons).

I find Fane's criticism of my account of why Keynesian policies in Britain seemed to work before 1967 puzzling. It is quite true that what he calls a "rational expectations model" would provide different results to the equation which I used. But that equation, mechanical as it may be, does at least have the support of some empirical evidence, being based on the work of Cross and the author. Fane's rational expectations framework, in which inflation expectations depended upon monetary expansion rates at home and abroad, and perhaps fiscal policy as well, has not, as far as one can tell from his paper, been put to any empirical test. It *may* provide a better explanation of the facts than does my equation, but it is a long step from asserting that it might do so to showing that it does. Moreover, I find it strange that Fane should espouse such a model, given his insistence that trade union militancy and rising import prices played an important role in generating British inflation. If union militancy and rising import prices *do* cause variations in the domestic inflation rate, then these variables, rather than monetary expansion rates and such, should be the determinants of rational expectations about the inflation rate.

But really, there is not all that much difference between Fane and myself about many of these matters. It is common ground between us that the natural unemployment rate increased after 1967 and that failure to recognize this fact lay at the root of the overexpansionary policies carried out in the 1970's. Reading between the lines of Fane's argument a little, I wonder if he would disagree that the exchange rate regime was important to the extent that under fixed rates such policies would have led to a balance-of-payments crisis—as they did in 1964—but that under a flexible rate their impact was diverted to the exchange rate and the price level. Certainly that would be my position. All that lies between us is a minor disagreement about whether the *average* rate of unemployment before 1967 was about right, as Fane suggests, or a little too high as I suggested. Here I would note that R. J. Ball and T. Burns appear to agree with me, and cite the undervaluation of sterling in the 1950's as a factor that prevented a major balance of payments problem materializing before the mid-1960's. That is a factor which I ought not to have neglected in my original paper.

Finally let me note one comment in Fane's paper with which I strongly agree. I do think it would be possible to reformulate at least some versions of the "union militancy" hypothesis of British inflation in natural unemployment rate terms, and I agree that if this were done some interesting testing of alternative points of view could be undertaken. Until such work is done, however, and shows that union militancy has made an important difference to the

natural unemployment rate in Britain since 1967, I remain impenitent about downplaying its significance.

REFERENCES

R. J. Ball and T. Burns, "The Inflationary Mechanism in the U.K. Economy," *Amer. Econ. Rev.*, Sept. 1976, *66*, 467–84.

J. A. Carlson and J. M. Parkin, "Inflation Expectations," *Economica*, Feb. 1975, *42*, 123–38.

R. J. Cross and D. Laidler, "Inflation, Excess Demand and Expectations in Fixed Exchange Rate Open Economies: Some Preliminary Empirical Results," in J. Michael Parkin and George Zis, eds., *Inflation in the World Economy*, Manchester 1976.

J. I. Foster, "The Relationship Between Unemployment and Vacancies in Great Britain 1958–72: Some Further Evidence," in David Laidler and David Purdy, eds., *Inflation and Labour Markets*, Manchester 1974.

D. Laidler, "Inflation in Britain: A Monetarist Perspective," *Amer. Econ. Rev.*, Sept. 1976, *66*, 485–500.

M. H. Miller, "Can a Rise in Import Prices be Inflationary and Deflationary? Economists and U.K. Inflation, 1973–74," *Amer. Econ. Rev.*, Sept. 1976, *66*, 501–19.

J. Taylor, "Incomes Policy, and the Structure of Unemployment and the Phillips Curve: The United Kingdom Experience 1953–70," in J. Michael Parkin and Michael T. Sumner, eds., *Incomes Policy and Inflation*, Manchester 1972.

House of Commons, *Ninth Report from the Expenditure Committee, Session 1974–75, Public Expenditure, Inflation and the Balance of Payments*, London 1974.

National Institute of Economic and Social Research, *Nat. Inst. Econ. Rev.*, Nov. 1973, No. 66.

Externalities, Extortion, and Efficiency: Comment

By Daniel W. Bromley*

In a recent article George Daly and J. Fred Giertz (D-G) assert that: "... general agreement has not been achieved regarding either the likelihood of extortion or its probable impact on the allocation and distribution of resources in environments characterized by harmful external effects" (p. 997). They proceed to define extortion as:

> ...the act of obtaining payments from some entity in return for *not* imposing upon that entity some harmful effect, where the generator of the external effect receives no direct net internal benefit from the act. It is very important to distinguish between what we have labeled extortion and other bargaining situations where the externality generating party does receive direct net benefits. [p. 998]

While this notion of extortion seems at odds with the conventional definition, the crucial weakness of the D-G formulation is the set of conclusions derived from their model; conclusions which appear to be artifacts of their theoretical construct.

They conclude that normal bargaining in situations where extortion might exist leads inexorably to maximum social welfare, while extortionistic bargaining could, under the most favorable circumstances (zero transaction costs), merely duplicate the results of independent behavior which produces allocative efficiency directly. They further conclude that (with nonzero transaction costs) extortion can only reduce social welfare. In the following I will demonstrate that: 1) normal bargining does not necessarily lead to a maximum social welfare; 2) even with zero transaction costs extortionistic bargaining need not merely duplicate the individual actions which would otherwise result; and 3) when transaction costs are positive extortionistic bargaining can yield enhanced social welfare.

I. On Property Rights, Pareto Optimality, and Public Policy

Before turning to an explicit treatment of extortionistic bargaining it is necessary to establish the relation between the assignment of property rights (entitlements) and my conclusions regarding Pareto optimality. Using E. J. Mishan's smoking example we can consider a two-person situation—smoker (S) and nonsmoker (N).

In Figure 1 the two points (L and $\bar{L}$) on the contract curve represent equilibrium positions of the two-person world under two possible structures of property rights. If smoking is allowed in public buildings we find ourselves at point L and individual N is on a lower indifference curve than if smoking were not allowed. Conversely, if smoking is not allowed ($\bar{L}$) then it is S who is dissatisfied. What L and $\bar{L}$ represent then, are two possible goods bundles, bundles which are made possible by a set of rules which govern behavior of the would-be smoker (S). Under either institutional structure we can assume efficiency in our two-person world.

If we assume a starting point an initial structure of entitlements—which is permissive of smoking in public places (call it L law) we start at point L in Figure 1; G shows the allocation of other goods as between S and N, while SS and SN show their respective consumption levels of smoke—a "good" for S, a "bad" for N.

If we take this entitlement as the status quo ante and inquire as to possibilities for negotiations between the two we might imagine that the issue would be a debate

*Professor, department of agricultural economics, University of Wisconsin-Madison. I am grateful to Andy Dragun, Basil Sharp, and several anonymous reviewers for valuable suggestions.

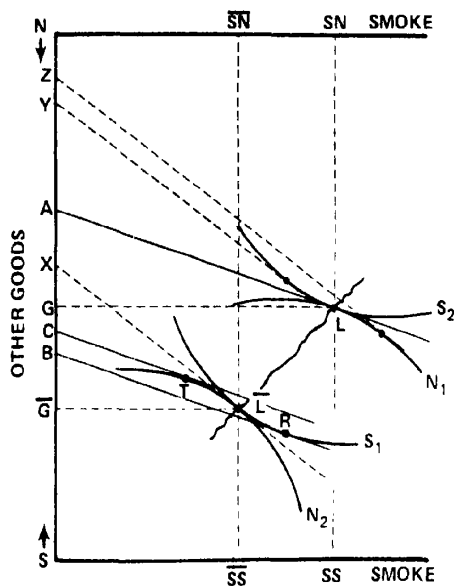



FIGURE 1

over the right (entitlement) to smoke in public places.[1] Notice here that the bargaining is not of the typical sort where efforts are directed toward a movement *to* the contract curve. Assume an initial situation in which all of those sorts of bargains have been exploited and the issue now is one of an abrupt change in public rules—a ban on smoking. This bargaining over property rights (or entitlements) has profound effects on the bundle of goods produced in society—we get a different output of trout and paper products depending upon whether it is the pulp mills or the trout lovers who possess the entitlement to streams and rivers; for it is the initial structure which then determines who must buy out whom.[2] Assume that N favors the prohibition of smoking (call it $\bar{L}$ law) and would be willing to pay a certain sum to S to agree to a change in entitlements.[3] The amount that N would be willing to pay to S to move from point L to $\bar{L}$ is the compensating variation and is reckoned in Figure 1 as the distance AC. However, in deciding whether or not to accept the offer S will compute the welfare loss as the distance AB; the net compensating variation is negative (B-C) and S would not agree to a change. An observer would therefore conclude that a law permitting smoking (L law) was optimal since it was impossible to improve N's utility without diminishing S's utility.

Now assume that we start from a structure of entitlements where smoking in public places is not allowed ($\bar{L}$ law). Here it is S who must bear the transaction costs of negotiating with N. The maximum offer which S would make to N is reckoned as XY, while N's loss from a change in entitlements is XZ. Again, the loss to the party opposed to the change exceeds the maximum willingness to pay of the proponent and N would not agree to the restructuring of entitlements. As before, the initial assignment of entitlements bestows an advantage on the already favored member of society and no "bargain" could be struck—the gainer (S) is unable to buy out the loser (N). An observer would therefore conclude that the law prohibiting smoking ($\bar{L}$ law) was optimal.

We thus see that the policy implications of costless (or nearly so) bargaining over the structure of entitlements are crucially dependent upon the initial assignment of rights. Under one initial condition a law permissive of smoking is considered "optimal," while under another initial condition a law prohititing smoking is considered optimal.

In a world with nonzero transaction costs, the incidence of such costs and the implications for the outcome are important. Under L law it is N who must bear these costs to induce a change in the law to $\bar{L}$; this further reduces the net gain (the compensating variation) to N and exacerbates the extent to which the current situation appears optimal. Under an $\bar{L}$ law where smoking is prohibited in public buildings the burden rests with S. Again,

[1] Indeed smoking is not permitted in an increasing number of public buildings (or portions thereof).

[2] For a more detailed discussion see the author.

[3] Assume a need for unanimity on such rule changes.

with nonzero transaction costs falling on S, the potential gain from a change in property rights is reduced and the current situation appears to be optimal. The apparent efficient (optimal) outcome is a function of the status quo distribution of property rights and income.

II. Enter Extortion

While extortion can be thought of as mere coercion, its more common connotation—and the one employed by D-G—carries the implication of payments made by a party to preclude something unwanted from taking place. There is yet another point which D-G chose not to stress—legal and illegal behavior. Perhaps the classic notion of extortion is the collection of "fire insurance premiums" from big-city businesses to prevent "accidental" fires.

In the current context, and under a permissive smoking law (L), extortion would exist only if N were able to extract a payment from S to permit smoking in public places. We might imagine that the price N would seek from S would be sufficient to compensate N because of the unwanted law—this would be the magnitude AC which would permit N to attain the same indifference curve as could be attained with $\bar{L}(N_2)$. Individual S would clearly gain by paying at least AC to N (assuming that N had the means to prevent smoking in the absence of the payment even though S had the entitlement), and might even be willing to pay (almost) AB; an extortionistic bargain could be struck, with N made at least as well off by the payment from S to smoke (and perhaps better off if part of AB were forthcoming), and with S still having a small surplus.

Whereas positive transaction costs merely reinforce the status quo ante in normal bargaining situations, here the outcome is significantly affected. The magnitude BC is the bargainable surplus and therefore sets the upper bound on the level of transaction costs; this is a function of the tastes and preferences of S and N. In the current example the transaction costs fall on N to extort a payment from S and we would therefore expect that very low transaction costs (including zero) would facilitate extortion. However, as such costs approach the magnitude of the bargainable surplus (BC) extortion would cease to be feasible since its costs would exceed the gain to the initiating party (N).

If we switch to an initial situation characterized by a ban on smoking ($\bar{L}$ law), extortion would enter only if S attempts to extract a payment from N in order to refrain from doing that (smoking) which is not permitted. The possible gain to S has a lower bound of XY; this is the amount that would leave S as well off as if smoking were permitted. S would try for more, but would settle for XY; N, on the other hand, would be willing to pay no more than XZ since this represents the loss if smoking were allowed. That is, N would surely be unwilling to pay more to prevent S from smoking since the former would be less well off than if S were allowed to smoke (L). Since XZ is the maximum that N would pay to prevent S from smoking in public buildings, it also represents the limit on bargaining before N would agree to vote for a change in the law (assuming unanimity for such changes). With zero transaction costs (TC) a bargainable surplus of YZ exists which would be sought by S. With S bearing the transaction costs, as such costs became large this would diminish the bargainable surplus until $TC = YZ$. Here, extortion would stop.

III. On Normal Bargaining, Extortion, and Social Welfare

Daly and Giertz conclude their analysis by stating that:

> Conventional bargaining ... can lead to a superior resource allocation and therefore, depending upon the size of the transactions costs, may improve social welfare according to the Pareto criteria. On the other hand, the possibility of ... (i.e., extortion) ... can under the most favorable conditions

> (with zero transactions costs) only duplicate the results of independent behavior which produces allocative efficiency directly. With any positive level of bargaining costs, extortion will clearly lead to a reduction of social welfare since scarce resources are utilized in the process of negotiation while failing to improve the allocation of resources. [pp. 999-1000]

Their jump from conclusions regarding efficient outcomes to those which are ideal in a social welfare sense is without analytical justification.[4] Every point along the contract curve in Figure 1 is efficient and (absent extortion) Pareto optimal. But moving from one Pareto optimal point to another Pareto optimal point—that is, moving from L to $\bar{L}$, or vice versa—may not be a Pareto improvement. Indeed, as seen here, in the absence of extortion it would not be an improvement.

In this section I will show that normal bargaining does not inexorably lead to a maximum social welfare; that requires a close correspondence between the social welfare function and the structure of entitlements (rights).

Additionally, it will be shown that even with zero transaction costs in extortionistic situations social welfare can be reduced. Finally, we will see that when transaction costs are positive, extortionistic bargaining can yield enhanced social welfare. To establish these conclusions, which are contrary to the D-G findings, it is necessary to consider two possible social welfare functions, one favorable to S, (W), and one favorable to N, $(\bar{W})$. In Figure 2 these two families of indifference curves are depicted with the utility-possibility frontier derived from Figure 1. There are four possible cases.

Case 1: *L law and* W_i *obtain*. With the family of social welfare curves and an as-

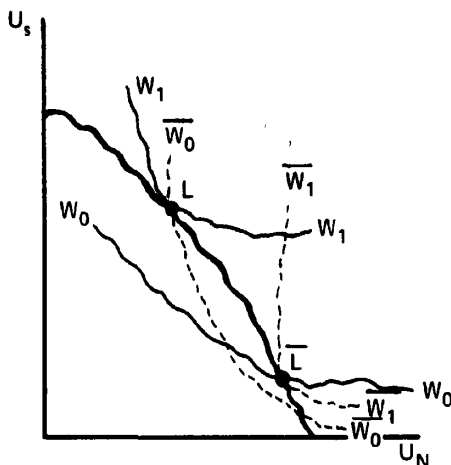



FIGURE 2

sumed starting point of a permissive law about smoking in public buildings (L) we are on W_1 at L in Figure 2. The earlier discussion indicated that no bargained solution would occur since the net compensating variation would have been negative under a change and hence we remain at L where social welfare is a maximum. Now consider extortion by N; when transaction costs are zero we would move toward the $\bar{L}$ position in Figure 1 but smoking would occur. This would move our two-person world to W_0 and social welfare has been diminished. When transaction costs are positive but less than BC we would also move close to $\bar{L}$ position but smoking would occur and social welfare (on W_0) has gone down. Only if transaction costs exceed BC would no extortion occur, in which case we would remain at L, smoking would occur, and social welfare appears to be a maximum.

Case 2: *L law and* $\bar{W}_i$ *obtain*. Now consider the same assignment of property rights (L) but a different family of social welfare curves $(\bar{W}_i)$. Attempts at normal bargaining as discussed earlier would result in no change in the situation, society would remain at position L in Figure 1, but this is on social indifference curve $\bar{W}_0$ and aggregate welfare is less than in the prior in-

[4]The frequent mixing of the terms *efficiency* and *maximum social welfare* in the D-G article indicates a special need here to be explicit in the context of extortion and externalities.

stance. Thus, the same assignment of property rights and efforts at a bargained solution can be seen to lead to two different conclusions about social welfare, depending upon our assumptions about the social welfare function.

When extortion is introduced the starting point is on $\overline{W}_0$ at L, and when transaction costs are zero we move to a position of $\bar{L}$ but smoking occurs. Such a move shifts us to $\overline{W}_1$ (point $\bar{L}$ in Figure 2) and social welfare has increased. When transaction costs are greater than zero, but less than BC, the same outcome results and once again social welfare has increased. Finally, when transaction costs exceed BC there is no change from the initial situation and social welfare remains at $\overline{W}_0$ (point L in Figure 2).

Case 3: *$\bar{L}$ law and $\overline{W}_1$ obtain.* Here, as seen before, normal bargaining would result in retention of $\bar{L}$ which implies position $\bar{L}$ on $\overline{W}_1$ in Figure 2; we would be led to believe that social welfare is a maximum. With extortion we have the same three possibilities. When transaction costs are zero we would observe a move toward position L in Figure 1 but there would be no smoking. We started out on $\overline{W}_1$ in Figure 2, but end up on $\overline{W}_0$; social welfare has fallen. When transaction costs are greater than zero but less than YZ we again would observe a move to position L but there would be no smoking. Again, the move from $\overline{W}_1$ to $\overline{W}_0$ decreases social welfare. Finally, when transaction costs exceed YZ we would not move from $\bar{L}$ and would remain on $\overline{W}_1$ in Figure 2; social welfare is a maximum.

Case 4: *$\bar{L}$ law and W_1 obtain.* In this case, normal bargaining would tend to keep us at position $\bar{L}$ in Figures 1 and 2, and we would be on W_0 in the latter. Again, attempts at normal bargaining would not result in any change, yet social welfare is less than it could be. With extortion and zero transaction costs we would move to position L in Figure 2 but there would be no smoking. We have moved from W_0 to W_1 and social welfare has increased. When transaction costs are greater than zero but less than YZ we would move to position L as before, and social welfare would increase. Finally, when transaction costs are greater than YZ we would stay at $\bar{L}$ on W_0 and social welfare stays below what it could be with extortionistic bargaining.

IV. Conclusions

In contrast to the D-G formulation of the economic implications of traditional bargaining vis-à-vis extortionistic bargaining, it is seen here that extortion *does not* merely duplicate the traditional bargaining situation. The magnitude and incidence of transaction costs as between the two cases are central to resource allocation, and the distribution of welfare between the two citizens. More importantly, extortion creates a situation in which nominally Pareto optimal outcomes—nominal because they are a function of the status quo configuration of property rights—can be judged by the two (or n) parties beyond the conventional bargaining model.

D-G conclude that the legal definition of extortion has a "rather firm economic basis which ... distinguishes it from other forms of bargaining. That is, the legal term extortion is applied to bargaining which can, like other forms of bargaining, redistribute wealth but which, unlike these forms, cannot improve the allocation of resources" (p. 1000). Yet we see above that extortionistic bargaining can indeed lead to improved welfare.

Clearly, gains from extortionistic bargains are present when laws (property rights assignments) are at odds with the prevailing social welfare function. In a society which declares that all individuals count equally, yet the laws permit, say, slavery, it is easy to imagine gainful extortionistic bargains where the slaves can extract some of the bargainable surplus and leave both parties better off then before. But the real issue is one of solutions which are presumptively Pareto optimal, and which seem to rule out extortion. While certainly not advocating this particular form of coercive bargaining, it is vital that we be clear about its impacts upon the "optimal" outcome,

and the attendant distributional implications.

REFERENCES

D. W. Bromley, "Property Rules, Liability Rules, and Environmental Economics," *J. Econ. Issues*, Mar. 1978, *12*, 43-60.

G. Daly and J. F. Giertz, "Externalities, Extortion, and Efficiency," *Amer. Econ. Rev.*, Dec. 1975, *65*, 997-1001.

E. J. Mishan, "The Postwar Literature on Externalities: An Interpretive Essay," *J. Econ. Lit.*, Mar. 1971, *9*, 1-28.

Externalities, Extortion, and Efficiency: Reply

By GEORGE DALY AND J. FRED GIERTZ*

Our analysis of extortion was undertaken in much the same spirit as that of many other recent authors who have written in the area of economics and law (for example, Ronald Coase, Harold Demsetz, Richard Posner, and Gordon Tullock). Until fairly recently, most of the interest shown by economists in the law was of two types: 1) the legal environment was accepted as given, and the economic effects of such an environment were studied with the tools of economic theory; or 2) certain goals were considered desirable and normative suggestions about changes in the legal environment necessary to achieve these goals were analyzed. The positive analysis of the law carried out in recent years has dealt with a different sort of question, that of what forces shape the legal environment. One answer to this question is that the legal system seems to respond (although obviously imperfectly) to considerations of economic efficiency.[1] In regard to the analysis of extortion, our goal was not the normative one of judging whether extortion should be encouraged or proscribed, but the positive one of explaining why extortionate types of bargains are made illegal by most societies while many other kinds of bargaining are encouraged by the law.

Our answer to this question is based upon the fundamental difference between the usual type of bargaining as analyzed by Coase and others, and that characterized by extortion. In conventional bargaining, each person gives up something of value in return for something valued even more highly, which results in a more efficient allocation of resources with both parties sharing the gains. In extortionate bargaining, however, there are no comparable efficiency gains. The victim of extortion is forced to compensate the perpetrator to refrain from doing something which does not directly benefit anyone and would not be undertaken save for its bribe generating potential. In a world of zero transactions cost, extortion would merely redistribute resources; in a world of costly transactions it would result in resources being used in the bargaining process with no resulting allocative gains, that is, it would result in a movement to a point further from the production-possibility frontier. It is our assertion that this, in part, explains the legal prohibition of extortion.[2]

Daniel Bromley's disagreement with our analysis stems from two major sources: 1) our ambiguous use of terminology which gives the impression that we have defined allocative efficiency and social welfare identically; and 2) a fundamental error in the model he presents. We will discuss these in turn.

*Institute for Defense Analyses and University of Houston, and Miami University (Ohio), respectively

[1]A rather extreme statement of such a position (which we do not fully subscribe to) is contained in Gary Becker: "... the methods used to accomplish any given end tend to be the most efficient available, in the public as well as the market sector.... Although this approach leaves little room for economists to suggest improved methods in the public sector, it gives them potentially a much enhanced role in the positive analysis of the laws of operation of this sector" (p. 248).

[2]It should be noted that the key feature of extortion is the use of a threat to elicit a payment, not whether the threatened action is legal or illegal An attempt to generate payments based upon a threat to do bodily harm or property damage is an example of extortion where both the threat and the threatened action are illegal. Another example of extortion is, however, blackmail where the threat involves the revelation of embarrassing information In this case, the threatened action of publicizing the information is perfectly legal in a free society, but the threat itself is still illegal. Why is the threat to do something which itself is within the law illegal in this case? We would suggest that in the context of our analysis, it is because the threatened action would yield no direct benefits to the blackmailer and would be undertaken only to elicit compensation

Bromley concludes that his analysis contradicts our alleged assertion that normal bargaining inexorably leads to maximum social welfare. He does this by demonstrating that a non-Pareto optimal point can yield a higher level of social welfare according to an externally imposed social welfare function than some Pareto optimal outcomes. We did not intend to imply that allocative efficiency and social welfare are identical and regret any confusion that our sloppy terminology may have generated. Our conclusion, correctly stated, is that in the presence of transactions costs, extortion leads to inefficiency, not necessarily reduced social welfare, and we are indebted to Bromley for clarifying this point.

This relatively simple clarification, however, does not require the analytical apparatus Bromley presents. Moreover, that apparatus makes a fundamental analytical error which is potentially far more confusing than any semantic problems in our paper.

In Sections I and II, Bromley's analysis concentrates on the possibility of bargaining among Pareto optimal points along the contract locus of a modified Edgeworth-Bowley box diagram.[3] Referring to his Figure 1, L represents the equilibrium achieved when the legal arrangements permit smoking. However, the nonsmoker (N) has the power to prevent the smoker (S) from exercising his legal right to smoke and can force the two-person society to point $\bar{L}$, another Pareto optimal point where smoking is at a lower level. Bromley suggests that a payment could be made by S to N in order to keep the society at L rather than move it to $\bar{L}$, that is, the smoker can pay the nonsmoker enough to induce him not to move along the contract curve in a direction favorable to the nonsmoker. Such a conclusion represents a bold challenge not only to our paper but also to the very concept of Pareto optimality since it implies that all parties can be made better off by a movement from one Pareto optimal point to another.

Closer examination, however, reveals that Bromley's analytics do not support his conclusion. He states that the minimum payment the nonsmoker would accept to refrain from moving to $\bar{L}$ is equal to AC. The use of the compensating variation technique assumes that N is free to use his payment to achieve a consumption point along CT at point T. (It is not clear what the slope of the price line represents in the context of this analysis. Is it the terms of trade between smoke and other goods established by some kind of market process? If such a price ratio is established, why do not both persons have to adjust to it simultaneously?) Point T would entail a level of smoke below that of point $\bar{L}$, and would place the nonsmoker, N, on the same indifference curve that passes through $\bar{L}$. Remembering that in a box diagram the final consumption point must be the same for both individuals, would the smoker agree to pay N's minimum price AC which would place the final consumption point for both at T rather than $\bar{L}$? The answer is obviously no since the smoker's indifference curve passing through T is below that of the one passing through point $\bar{L}$. Clearly, no bargain could be struck.

Bromley goes on to suggest that the smoker would pay a maximum of AB to avoid being forced to consume at point $\bar{L}$. In computing the compensating variation, it is again assumed that S is free to move along the constraint BR until a maximum is achieved at R. From this analysis, it is clear that S could pay up to AB and be no worse off than being at point $\bar{L}$. Again, the question must be asked if N would accept point R (both must again consume at the same point) rather than exercise his power to move to $\bar{L}$? Nonsmoker N would clearly reject such a deal since consumption at point

[3] Bromley's analysis deals with a different bargaining situation from the one we analyzed. The key feature we emphasized, which usually characterizes extortion, is the absence of any direct net benefits to the perpetrator from the threatened action used to elicit the payment. Bromley, however, focuses on a quite different bargaining framework where the threatened action involves important direct net benefits to the extortionist, in that the threat involves taking resources (or rights) from the victim for the extortionist's personal use.

? puts him on a lower indifference curve han the one passing through point $\bar{L}$.[4]

It is clear that the compensating variaion technique as used by Bromley is in:orrect and leads to the mistaken concluion that: "... extortion creates a ituation in which nominally Pareto opimal outcomes—nominal because they are ι function of the status quo configuration of property rights—can be judged by the wo (or n) parties beyond the conventional bargaining model" (p. 734). When the inalytical devices employed by Bromley are ised correctly (i.e., compensation payments ire actually made and both individuals nust end up at the same point in the box liagram), we can see that what we have nown all along is true: there is no way S an pay N not to move costlessly along the ontract curve in a direction favorable to V. A similar sort of analysis would also lemonstrate the same error in Bromley's rgument when property rights are reversed nd smoking is prohibited at $\bar{L}$. If the moker has the power to violate the law nd move costlessly to L, there is no way hat the nonsmoker can pay him not to do o.

In the absence of transactions costs, no xtortionate payment could be made to revent a movement along the contract urve. With various kinds of transactions osts, however, extortion (at least in a tatic situation) might be productively analyzed in the framework used by Bromley. It is likely that an individual could move along the contract curve only with the outlay of resources. These costs may involve lobbying costs to change the legal environment or they may entail resources devoted to illegitimate activities to appropriate the gain. Likewise, the other person will be prepared to devote resources to protect his favored position. In such a situation, extortion might lead to a preferred position where these costs could be minimized. These costs, unfortunately, were not introduced in Bromley's work.[5]

[4]The futility of attempting to pay N not to move ong the contract curve can be viewed in another way s well. If N has the power to move to $\bar{L}$, the nonnoker would accept no outcome resulting in a level of tility less than N_2. Given this constraint, the high-t level of utility the smoker can achieve is S_1 at point . Since the outcome $\bar{L}$ is the very best the smoker in achieve given the nonsmoker's power, there is no dvantage to bargaining between N and S to move way from $\bar{L}$.

[5]We are puzzled by Bromley's analysis of bargaining with and without transactions costs in Section III. Positive transactions costs do not seem to be reflected in the diagrammatical analysis. Presumably, the resources used in bargaining would move the final outcome somewhere inside the original utility-possibility frontier

REFERENCES

G. Becker, "Comment," *J. Law Econ.*, Aug. 1976, *19*, 245-48.

D. Bromley, "Externalities, Extortion, and Efficiency: Comment," *Amer. Econ. Rev.*, Sept. 1978, *68*, 730-35.

R. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1-44.

G. Daly and J. F. Giertz, "Externalities, Extortion and Efficiency," *Amer. Econ. Rev.*, Dec. 1975, *65*, 997-1001.

H. Demsetz, "Towards a Theory of Property Rights," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 347-59.

Richard Posner, *Economic Analysis of Law*, Boston 1973.

G. Tullock, "Welfare Costs of Tariffs, Monopolies and Theft," *Western Econ. J.*, June 1967, *4*, 224-32.

IN MEMORIAM
HARRY G. JOHNSON
1923–77

Remarks of Arnold C. Harberger at the Presentation of AEA Distinguished Fellow Award to Harry G. Johnson, December 29, 1977

The letters that flooded in from around the world in the weeks and months following Harry's death are ample testimony to the great shock, sorrow, and grief that the economics profession felt at his loss. I cannot here record the many themes that were touched in these letters, but again and again they said that Harry was an institution, that he was larger than life, that he somehow seemed to have fitted four or five lifetimes of work and experience and influence into a bit under 54 years. These letters, many of them from people unknown to me, came from the heart. They bespoke an appreciation that is rarely accorded by a profession to one of its members.

It is not surprising, to those who knew him, that Harry meant so much to the profession, for there have been few economists in this or any other era who gave so much of themselves to it. No single element pervades the many facets of the life of this very complex, wonderful, and great man as much as his relations to and interactions with the profession of economists. This is why I have chosen to offer in tribute to Harry what I believe was his own code of professional conduct.

The professional economist has a duty:

1) To work so as to continually expand and update the insights that economic science can offer. (Here were Harry's scientific contributions.)

2) To draw the lessons that flow from an understanding of economics and its history for the formulation of policy. (Here were Harry's many policy papers.)

3) To expose frankly and unequivocally, in public forums, the fallacies that lie at the root of common beliefs and judgments, and of many political platforms, concerning economic policy. (Let us not forget, Harry's was the most cogent voice in the entire debate surrounding nationalist economic policy in Canada. He was also the consistent debunker of myths about the process of economic development, where he had strong and sobering messages for aid-givers and aid-receivers alike.)

4) To be forthright and unstinting in criticizing the work of his professional peers (the idea of the profession as a club, in which criticism among members should be muted if it was expressed at all, was absolutely anathema to Harry).

5) To be wrathful in exposing unworthy behavior by fellow professionals. (Harry's wrath fell mainly on some of the mighty within the profession, when he viewed them as misusing the authority they had gained on the basis of their legitimate past contributions. For Harry, an even greater burden of professional responsibility fell on an individual, once he had gained such authority.)

6) To heed, insofar as is humanly possible, calls to share his professional knowledge with others. (Harry's travels on the lecture circuit were legendary; it would be hard to calculate the amount of influence he exercised in this way.)

7) To help, above all, the young people who are struggling to expand their knowledge and to grow within the profession. (This was Harry's real soft spot. For all the renown he achieved for sharp criticism and barbed response, it is amazing how the younger and weaker members of the profession escaped. Obviously this was no chance event, it was Harry's policy. As editor, he could be ruthless in rejecting the work of a senior colleague, yet he spent countless hours drafting careful suggestions to young unknown authors. When he went away to lecture at other universities and other countries he used to come back with little sheets of paper, scribbled with notes to himself as to the writings he had prom-

ised to send one man, the interview he had promised to arrange for another, the manuscript he had agreed to review for a third.)

8) To keep his promises, always. (This may be the most incredible aspect of this incredible man. In all the twenty-odd years that I knew Harry, I'm not aware of a single instance in which he failed to perform as promised. This was true with respect to big things, like writing promised papers on time or somehow adjusting an impossible schedule so as to accommodate yet more commitments without failing on any of the previous ones. But it was also true with respect to little things, like reading student papers as promised, turning examination grades in on time, performing responsibly on committee assignments, etc. Sometimes I really felt that Harry was at one and the same time the man with the best excuse for not doing small departmental tasks, yet the one who accepted them most willingly, and carried them out in the most timely and responsible way.)

9) To avoid arrogance and pomposity at all costs. (Humility is not part of Harry's image, within or without the profession, yet I think that he was fundamentally a humble man in the best and most honorable sense of that word. For all his achievements, he was always ready to do his part in giving an extra lecture, writing yet another paper, traveling to some far corner of the world to preach the gospel of good economics there. He was the farthest thing from a prima donna, never asking special considerations for himself, always performing willingly tasks that brought complaints from at least some others.)

10) To earn his place in the profession, every day. (Harry was as strong in refusing to rest on his laurels as he was in reacting against others resting on theirs. Harry was a master craftsman, and like most true craftsmen, he was not content to be idle. He spent little time in retrospective looks at his past labors, but instead devoted his energies to moving ahead. He never wanted to be known for how good he had been in the past, always for how good he was right now.)

Well, these are Harry's precepts for professional life, as best as one of his old friends can infer them from his words and actions over the years. I believe they are accurate in representing the principles Harry believed in, judged others by, and, most importantly, lived by himself. I believe, too, that our profession would be stronger and better, more productive scientifically, and more socially useful, if the rest of us came closer than we have up to now to meeting these standards of professional behavior.

Arnold C. Harberger
The University of Chicago

NOTES

The 1978 Employment Center will be held December 28–30 at the Conrad Hilton in Chicago, Illinois. Operating hours will be December 28, 10:00 A.M.–5:00 P.M., December 29th and 30th, 9:00 A.M.–5:00 P.M. Special room rates have been established for the meeting ($25/singles, doubles and twins, $30/deluxe twins, $50/one-bedroom suites and $75/two-bedroom suites.) Hotel reservation cards will be mailed to you upon receipt of your placement form.

Requests for placement forms should be made to Ms. Kathy Nichols, National Registry for Economists, Illinois State Employment Service, 40 West Adams Street, Chicago, IL 60603 between September 1 and November 15. *Note:* No forms will be mailed prior to September, however they will be available at the American Economic Association booth at the annual meeting. Completed forms *must* be returned by December 1. You do not have to attend the meeting to have your listing on file. There is no registration fee.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

Call for Papers. The Eastern Finance Association will hold its annual meeting in Washington, D.C., April 19–20, 1979. There will be papers and discussions by academics, by business professionals, and by government specialists on almost all aspects of domestic and international finance. If you wish to participate, send a two-page abstract of your proposal before November 27, 1978 to Professor Michael Keenan, Vice President Program, New York University GBA, 100 Trinity Place, New York, NY 10006.

The Center for Population Research is emphasizing the need for increasing the amount of behavioral-social science research on population, family planning, and reproductive behavior. Social scientists are invited to submit research grant proposals, which upon receipt will receive careful evaluation through the usual peer review procedures. Attention is also directed toward the postdoctoral fellowship as a mechanism for developing careers in the population field. Interested persons are encouraged to apply, and applications will be evaluated by peer reviews. The deadlines for submission of research grant proposals are March 1, July 1, and November 1, while deadlines for postdoctoral fellowships are February 1, June 1, and October 1. Proposals are submitted on appropriate forms which may be obtained at most colleges and universities throughout the country. If information is needed, or to discuss any matter, please call 301+496-6515, or write Population and Reproduction Grants Branch, Center for Population Research, National Institute of Child Health and Human Development, Landow Bldg, Rm C-733, Bethesda, MD 20014.

The annual meeting of the Association for Economics and Social Science (Gesellschaft für Wirtschafts- und Sozialwissenschaften-Verein für Socialpolitik) will be held September 24–26, 1979 in Mannheim. The theme will be "Depletable Resources" and the following problem areas will be discussed: Optimal intertemporal allocation of renewable and nonrenewable resources, Allocation through markets and nonmarkets, International aspects of resource allocation. Proposals for papers are called for. Each one of the three areas will be split into several working sessions, with 3–4 papers per session. A committee will review the proposals which should be sent to: Professor Dr. Horst Siebert, Lehrstuhl für Volkswirtschaftslehre und Aussenwirtschaft, Universität Mannheim, 6800 Mannheim.

The P. W. S. Andrew Memorial Prize is awarded annually for an essay by a young scholar (under the age of 30 or within 8 years of taking his first degree) in the general field of Industrial Economics and the Theory of the Firm, broadly interpreted. The prize is £200 (or the equivalent in other currency) and the winning essay

will normally be published in *The Journal of Industrial Economics* An essay submitted should be a work of original research by the candidate only, not previously published and not previously awarded any other prize. It should be submitted in English and should not normally exceed 10,000 words in length. The closing date for entries is December 31 in each year Intending candidates for the prize should obtain details of the conditions of entry from the Administrative Officer, Office for Student and College Affairs, University House, University of Lancaster, Lancater, LA1 4YW, England. The winner of the prize for 1977 was D W Carlton, University of Chicago

---

Retirement

Ralston D. Scott professor emeritus of management, Southern Illinois University-Edwardsville, Aug 31, 1978

Promotions

William J. Boyes associate professor of economics, Arizona State University, Aug 1978

Thomas M. Carroll. associate professor of economics, Memphis State University, Sept. 1978

Roger K. Chisholm professor of economics, Memphis State University, Sept. 1978

Richard R Cornwall· associate professor of economics, Middlebury College

Alan C. DeSerpa associate professor of economics, Arizona State University, Aug 1978

K. J. Fung: associate professor of economics, Memphis State University, Sept. 1978

Elmer R. Gooding professor of economics, Arizona State University, Aug 1978

Andrew Hau· professor of economics, Millersville State College, Sept. 1, 1978.

Wolfgang Mayer· professor of economics, University of Cincinnati, Sept 1, 1978

Daniel A. Pavsek. assistant professor, department of economics, Baldwin-Wallace College, Sept 1, 1978.

Dominick Salvatore: professor, department of economics, Fordham University, Sept 1, 1978

Mark Satterthwaite· professor, managerial economics and decision sciences, Northwestern University, Sept. 1, 1978.

Patrick J. Welch: associate professor, department of economics, St. Louis University, July 1, 1978.

Administrative Appointment

Abby J. Cohen assistant vice president, T. Rowe Price Associates, Inc., Baltimore, Apr. 10, 1978

New Appointments

Josef C. Brada, New York University. associate professor, Arizona State University, Aug. 1978

Darius J Conger, Central Michigan University· visiting assistant professor, Arizona State University, Aug. 1978.

Judith M. Considine. instructor, department of economics, Rutgers College, July 1, 1978.

Antonio-Gabriel Cunha, Queen's University: economist, finance section, Office of Fiscal Analysis, Connecticut General Assembly. Dec 2, 1977

Anthony C. Fisher, University of Maryland· professor of energy and resources, and of economics, University of California-Berkeley, July 1978

Wallace C. Hardie instructor, department of economics, Iowa State University, Dec. 1, 1977

Robert J Hauser· research associate, department of economics, Iowa State University, Dec 1, 1977

J Kent Hill, Rice University· assistant professor of economics, Arizona State University, Aug. 1978

Ralph H Kline assistant professor, department of economics, Rutgers College, July 1, 1978

Thomas A Layman, University of North Carolina assistant professor of economics, Arizona State University, Aug 1978

John M McDowell, University of California-Los Angeles assistant professor of economics, Arizona State University

Paul Milgrom, Stanford University assistant professor, managerial economics and decision sciences, Northwestern University, Jan. 1, 1979.

Robert A Moffitt. assistant professor, department of economics, Rutgers College, July 1, 1978.

Bruce M. Owen, Stanford University associate professor of business administration and director of the Center for the Study of Regulation of Private Enterprise, Duke University, Sept. 1978.

Meir Statman instructor, department of economics, Rutgers College, July 1, 1978.

Nancy Stokey, Harvard University. assistant professor, managerial economics and decision sciences, Northwestern University, Sept. 1, 1978.

Resignation

Nancy Dayton Sidhu, Northeastern Illinois University Commercial Research Division, Inland Steel Company, June 1977

THE AMERICAN ECONOMIC REVIEW

December 1978

VOLUME 68, NUMBER 5

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this *REVIEW* is assumed by the editors or the publishers, The American Economic Association.

Articles

Shorter Papers

A Model of the Southern African-Type Economy

By RICHARD C. PORTER*

The purpose of this article is to give a broad, stylized picture of how Southern African economies "work," of the behavior of the economic actors, of the constraints to and goals of white policy, and of the conflicts and inefficiencies of resource allocation. The model is heuristic, that is, aimed primarily at understanding rather than empirical application. It is sufficiently removed from an exact replica of the South African or Rhodesian economy that it is more appropriately labeled "the Southern African-type" economy.[1]

The basis of the Southern African-type economy is a market economy where market constraints and policy parameters are determined by whites and for whites. Despite this dominance, there are many restrictions on the feasible range of white policy, and there are fundamental conflicts between different white groups and different white goals. Despite near complete power to fix white wage rates well above black wage rates and to preclude employers from hiring blacks to replace more costly whites, white policymakers cannot fully exercise their power lest they generate politically unacceptable levels of white unemployment (see Section II). Even when full employment of white workers is achieved, the resource allocation is economically inefficient—that is, the maximum potential white income is not realized. Further, there are conflicts between the interests of white capital and white labor; and the goal of high white income is in conflict with other white goals, namely, "industrialization" and reduced economic dependence on blacks (see Section III).

Over time, if the capital stock grows more rapidly than the white labor force, these conflicts are intensified between white capital and labor and between the other white goals of growth, namely, industrialization and reduced dependence on blacks (see Section IV). Finally, the sense in which whites "exploit" blacks is explored: essentially it is that the potential gains of integrating the capital-abundant white economy with the labor-abundant black economy are captured by the whites (see Section V).

I. The Model

To understand the basis of the Southern African-type economy, it is useful to consider three sectors, one where black labor works without capital, a second where black labor works with capital, and a third where black and white labor work together with capital. The three sectors reflect in a simple way the actual spectrum of black-white, labor-capital relations.

1) *Reserves*: In this sector, black labor works alone to produce output with constant average productivity of labor:

$$X_R = bL_R^B \tag{1}$$

where X is output, L is labor input, the subscript R refers to the sector (reserves), and the superscript B to the color of the worker (black). The simplifications implicit in this production relation need some defense. In the reserves of South Africa—also called "homelands" or "Bantustans"—the agriculture is tribal, communal, traditional, and extensive. Thus, while they are hardly devoid of land and capital, the capital is miniscule and largely self-produced, labor and land are applied in fairly fixed proportions, and constant returns to labor is probably a

*University of Michigan. I am particularly indebted, for criticism of earlier drafts, to G. H. Borts, A. V. Deardorff, S. L. Engerman, J. B. Knight, M. T. Nziramasanga, T. E. Weisskopf, and the referee. A longer version of this paper is available (Center for Research on Economic Development, disc paper no. 60) in which the South African foundations of the model and the development of the theory are handled more elaborately.

[1] For earlier models, see Stephen Enke and J. B. Knight.

satisfactory approximation. These reserves play three roles: i) the standard of living there, b per worker, provides a floor on which the black wage is elsewhere based;[2] ii) the reserves offer unlimited supplies of unskilled black labor to growing sectors; and iii) they provide a functional location for all black labor not demanded elsewhere in the economy.[3]

2) *Agriculture*: The essence of this sector is that its capital is white owned and its labor is black:

$$X_A = X_A(K_A, L_A^B) \tag{2}$$

where the subscript A refers to agriculture, the variable K represents capital, and the function, $X_A(\cdot)$ displays constant returns to scale and diminishing returns to each factor. Two critical simplifications should be noticed. One, there is no sector of the Rhodesian or South African economy in which white labor is not in fact found, at least in a supervisory or managerial capacity. In agriculture and mining, however, white labor is of trivial quantitative importance. And two, land and any concomitant diminishing returns to it are neglected. Despite the large size of Southern Africa—in terms of cultivable nonreserve area per rural worker—defense of this assumption really rests on the grounds that the insight lost is small in comparison to the additional complexity incurred by explicit consideration of land.

3) *Industry*: Both black and white labor work with white-owned capital to produce output in this sector:

$$X_I = X_I(K_I, L_I^B + L_I^E) \tag{3}$$

where the subscript I refers to industry, the superscript E refers to "Europeans,"[4] and the function $X_I(\cdot)$ displays constant returns to scale and diminishing returns to each factor. Here, as with the other two sectors, the model presents a greatly simplified stereotype of reality. To begin with, one should note that white labor is quantitatively important not only in "industry" but also in a variety of public utility, commerce, and service (including government) sectors. Most critical, and warranting careful examination, is the assumption that white and black labor services are perfect substitutes. Indeed, that black and white laborers are not treated as perfect substitutes is the very essence of the Southern African-type economy. But the equally clear concomitant is that the differential treatment is not justified by innate productivity differences; and the simplest way to capture this is to assume a one kind of labor model in which black and white labor are productively identical.[5]

Racial discrimination has appeared in many forms in South Africa—through access to education, apprenticeship, or on-the-job training, through access to certain occupations, through white-black employment ratios, through union contracts, through direct government prohibitions, penalties, or rewards, through informal pressures, and through cultural predilection.[6] Empirically, the most important means of discrimination today is the first of the above (see Knight and Michael McGrath); black workers simply cannot acquire the education and training necessary to qualify for the more skilled and better paid jobs. Here, in a model with only one kind of labor, no unions, and no education or training, it is helpful to consider an historically, though not currently, more important technique of discrimination, the job-reservation ratio,

[2]The rural-urban migration decision is greatly simplified by assuming the marginal and average product to be equal (to b) and by ignoring migrant labor and the structure of the black family. See Walter Elkan.

[3]Not incidentally in South Africa, the reserves also provide a separate *geographic* location.

[4]Labels are, I realize, fraught with values, but here I choose "Europeans" over the more logical "whites" simply to save the letter w for wage rates. Note that, throughout, the additional and differential South African color bars facing "colored" and "Asian" workers are ignored in this simple stylization.

[5]More realistic, but more complex, would be a two kinds of labor model with skilled and unskilled labor and capital all being imperfect substitutes (see P. R. Fallon and P. R. G. Layard). The discrimination then derives from the process by which white labor becomes skilled and black labor remains unskilled. The greater realism of such a model is probably not worth the price in terms of greater analytical complexity; nevertheless, some ideas about that model are offered in footnotes throughout the paper.

[6]For a history and description of the many facets of discrimination, see G. V. Doxey.

whereby a certain fraction of each employer's workers must be white (i.e., European). Thus,

$$L_I^E = c(L_I^B + L_I^E) \tag{4}$$

where c is the fraction of the total employment in the industrial sector reserved for European labor.[7]

It is assumed that all actors in the agricultural and industrial sectors act competitively in both product and factor markets. Further, the economy is seen as "small and open," which means that world prices are unaffected by the supplies and demands of this economy. Thus we can take internal prices as determined completely by external market conditions and internal policy decisions—that is, prices are exogenous to the model. (For convenience, all output units are normalized so that the price of each physical unit is one monetary unit.) Demands for products can be ignored since any sector's excess demand or supply can always be removed through international trade at the given and exogenous world market price.[8] The various assumptions and omissions are made partly for simplicity, but mostly because it seems important to show that neither skill differentials nor the government budget nor external trade nor monopoly power is essential to an understanding of the allocative and distributional workings of such an economy.

Competitive profit-maximizing producers in the agricultural and industrial sectors employ labor up to the point where its marginal revenue product equals its wage. In agriculture, where only black workers are hired, this means simply that

$$\delta X_A/\delta L_A^B = w^B \tag{5}$$

where w^B is the wage rate for black labor and δ refers to the partial derivative. In industry, the same criterion applies to the labor hiring decision, but the wage rate is more complicated; since hiring one worker means hiring a fraction c of whites and a fraction $1 - c$ of blacks, the relevant wage rate of one worker is a weighted average of the two wage rates:

$$\delta X_I/\delta(L_I^B + L_I^E) = cw^E + (1 - c)w^B \tag{6}$$

where w^E is the wage rate of European labor.[9] I assume that, through government, management, and union actions, the three parameters (c, w^B, and w^E) are exogenously specified—although we shall see in later sections that not all combinations of c, w^B, and w^E are feasible.[10]

Profit-maximizing firms also allocate capital so as to attain equality between its marginal revenue product and its cost:

$$\delta X_A/\delta K_A = r_A \tag{7}$$

and

$$\delta X_I/\delta K_I = r_I \tag{8}$$

where r_A and r_I are the rates of return to capital in agriculture and industry, respectively. The total capital stock ($\overline{K}$) is deployed between the two capital-using sectors,

$$\overline{K} = K_I + K_A \tag{9}$$

according to a function of the relative rates of return in the two sectors:

$$K_A/K_I = k(r_A/r_I) \tag{10}$$

where $k' \geq 0$.[11] If capital markets were perfect, then equation (10) would become $r_A =$

[7] Sally Frankel has called this the "multi-racial team system": "Over large sections of economic enterprise those responsible can increase or decrease the size of the team, but they cannot easily vary its proportionate racial composition. ." (p. 120). Were two kinds of labor considered, c might be determined technologically if skilled (European) and unskilled (black) labor had to be used in fixed proportions.

[8] If internal prices are different from world prices, such trade will generate government revenues or expenditures. I ignore these for simplicity, although the model would gain a giant step on reality if the government budget policies were considered.

[9] For simplicity I assume that $w_A^B = w_I^B$, although that is neither necessary nor realistic in the Southern African context. Implicitly, I think of w^B as greater than b—which is realistic, though also not necessary. In short, the sectoral mobility of black labor is sufficiently restricted through "influx control" that sizeable intersectoral black wage rate differentials can be, and have been, maintained. Needless to say, I consider only situations where $w^E > w^B$.

[10] In a two kinds of labor model, with substitution possible between skilled (European) and unskilled (black) labor, technology plus the wage rates w^E and w^B might determine c, the proportions in which black and European labor are employed.

[11] The prime refers to the derivative of the function $k(\cdot)$.

r_I. Finite values of k' are today more realistic in the Southern African context, and the extreme of sector-specific capital, that is, k' equal to zero, will occasionally be considered.

Finally, both black and white labor forces must be accounted for. Whatever black labor is not demanded by agriculture and industry is sent to (or more accurately, not permitted to leave) the reserves, so that

$$\overline{L}^B = L_A^B + L_I^B + L_R^B \tag{11}$$

where $\overline{L}^B$ is the total (exogenous) black labor force. Public policy in Southern Africa has been traditionally and strongly intolerant of white unemployment, so that for political equilibrium the system requires

$$\overline{L}^E = L_I^E \tag{12}$$

where $\overline{L}^E$ is the total (exogenous) white labor force.

II. The Solution

Since the system of twelve equations which comprise this model is largely recursive, it is possible to solve it sequentially, and in the process gain an understanding of the underlying economic mechanism. Consider equations (2), (5), and (7). With constant returns to scale and diminishing marginal products, the black wage rate determines output per worker, capital per worker, and the rate of return to capital in agriculture. Similarly for the industrial sector- equations (3), (6), and (8). Given the wage rate (i.e., the average of the wage rates of white and black labor, weighted by the fractions in which they are employed), output per worker, capital per worker, and the rate of return to capital are determined.

The total stock of capital at any time will be allocated between the two capital-using sectors (equation (9)) according to the relative rates of return to capital earned in these sectors (equation (10)). Since these rates of return are already determined by the production functions once wages are set, they are inalterable despite the mobility of capital; preferences of investors, given these rates of return, then determine the absolute size of the capital stock in each sector. Imperfect mobility of capital would insure that both sectors would exist even when $r_A \neq r_I$. Once capital is allocated between the two sectors, the prior determination of factor proportions means that the absolute level of output and employment in each sector is determined. Thus for agriculture L_A^B is determined; for industry the sum $L_I^B + L_I^E$ is determined. Then job-reservation rules (equation (4)) determine the racial composition of the industrial work force. Once demands for black labor are satisfied in agriculture and industry, the remaining black workers are "allocated" to the reserves (equation (11)).

The economy has allocated its resources. There is, however, no reason to suppose that the white labor demanded by industry will be equal to the white labor force (equation (12)).[12] Should unemployment appear in their ranks, whites can act to alleviate it through alterations in one (or more) of the key parameters of the model, c, w^E, and w^B. After all, these are not fixed by technology but rather by (white) policy; hence they are subject to change through union-management negotiations and/or through government minimum wage and job discrimination policies. Let us consider the impact on white employment of a change of c, w^E, and w^B, each separately.[13]

1) *Change in the white wage rate* w^E: A reduction in the white wage rate lowers the average rate in industry and hence reduces the capital-labor ratio and increases the rate of return to capital there. This draws at

[12] Mathematically, there are twelve equations, eleven variables, and parameters. The solution requires that at least one of those parameters be treated as a variable.

[13] Throughout, we will treat these parameters as independent and under the control of "policy." This is of course solely for analytical convenience; there are innumerable political and historical forces pushing, constraining, and linking these policy parameters. Indeed, most of the economic writing on South Africa is concerned with these forces—and more specifically, with the question of whether growth and industrialization tend to end or to perpetuate discrimination. For a summary of this debate, see David Yudelman.

least some new capital into the industrial sector. The total labor force in the industrial sector is increased in two ways, by the increase in K_I and by the fall in $(K/L)_I$. Since white labor makes up an unchanged constant fraction (c) of the total industrial labor, white and black employment in industry both rise. Cuts in the white wage rate can therefore serve to ameliorate white unemployment.

2) *Change in the job-reservation ratio c*: An increase in the job-reservation ratio means a higher effective wage rate in industry, which raises its capital-labor ratio, lowers its rate of return to capital, and induces a movement of capital out of the sector. The total labor force in the industrial sector is reduced in two ways, by the decline in K_I and by the rise in $(K/L)_I$. What happens to white employment is not clear—whites form a larger fraction of industrial employment, but total industrial employment has fallen. Thus, an increase in the job-reservation ratio (c) is not a sure cure for white unemployment; indeed, a decrease in c may be called for to increase white employment. Whether white employment rises or falls when c increases depends primarily on the degree of convexity of the industrial production function.

3) *Change in the black wage rate* w^B: A reduction of the black wage rate would appear to be the surest means to white full employment—as well as the politically most acceptable means (to whites).[14] Such a reduction reduces the effective wage rate in the industrial sector and raises the rate of return to capital there. This in turn reduces the capital-labor ratio and draws capital into the sector, both forces for increased employment and hence, given the job-reservation ratio, increased white employment.

There is, however, one problem with this reasoning. The lower black wage rate applies to the agricultural sector as well; thus the rate of return to capital increases there too. Whether capital moves into or out of the industrial sector depends upon whether the rate of return to capital rises more in industry or agriculture. Accordingly, the reduction of the black wage rate is not a certain means to increase white employment; an increase in black wages may be needed to increase white employment. Of course, if the mobility of black labor between agriculture and industry could be sufficiently restricted, it might be possible to reduce the black wage rate in industry but not in agriculture, which would have the desired effect on white industrial employment.

In sum, the model does not yield a solution for just any values of the key policy parameters: the black wage rate w^B, the white wage rate w^E, and the job-reservation ratio c. Despite their control of the political mechanism, and their near complete power to determine the mobility, job opportunities, and wage rates of blacks, whites are constrained by economic reality. As long as whites are unwilling to accept unemployment as part of the solution, there are limits to the values of w^B, w^E, and c they can select.

III. Conflicts

Recall that, by the Southern African-type economy, I mean an economy where the whites use their position of political dominance to constrain the opportunities of blacks in order to raise white incomes. Yet, even if the white population were so monolithic as to seek no goal other than the maximization of its own total income,[15] the process is not easy from a political point of view. There is an easy part: the black wage must be put as low as possible, consistent with the ability of agriculture and industry to attract the black workers they need from the reserves. Then comes the hard part. Maximization of total white incomes implies an economically efficient solution; and efficiency implies in turn that the private cost of labor to capitalists equal the social opportunity cost. This means that all industrial and agricultural labor, white as well

[14] Provided, of course, that the wage rate remains sufficiently above the black worker's opportunity cost in the reserves that the reduction does not dry up the flow of black labor to agriculture and industry.

[15] That is, the sum of $w^E L_I^E$, $r_I K_I$, and $r_A K_A$.

as black, should be priced at the same rate, namely the constant marginal value product on the reserves (b) plus (or minus) whatever differential is required to induce blacks to leave the reserves.[16]

In short, efficiency requires that identical factors of production be priced identically. But politically, in the Southern African-type economy, this is impossible. The mechanism which would be required, namely, taxation of high returns to white capital in order to make transfers to low-wage white laborers, neither exists nor is thinkable.[17] Indeed, one of the oldest and strongest foundations of South African economic policy has been its "civilized labor" policy, whereby the white wage rate must always be high enough to maintain, without income supplement, "the standard recognized as tolerable from the usual European standpoint."[18] Thus, with profit-maximizing competitive capitalists and without a system of transfers from white capital to white labor, an efficient solution is not practicable.

Where there is no acceptable mechanism for redistributing income between white capitalists and white workers, the tradeoff between the two white shares becomes an allocative as well as a distributive problem of public policy.[19] To illustrate this tradeoff, I make two further assumptions, that capital is completely immobile between sectors and that the black wage rate is already or elsewhere determined. Then the rate of return to capital in agriculture is determined and hence also the total earnings of white capitalists in agriculture. The problem then reduces to that of finding the income possibility frontier between white labor income and white capital income within the industrial sector. Formally, policy seeks to maximize white labor incomes in industry ($cw^E L_I$) subject to three constraints:[20] i) a given level of white capital income (i.e., $r_I \overline{K}_I$ = a constant); ii) full employment of white labor (i.e., $cL_I = \overline{L}^E$); and iii) marginal product determination of labor hiring (i.e., $\delta x_I/\delta L_I = cw^E + (1 - c)\overline{w}^B$). But the constraints leave nothing to maximize. The first constraint, the floor to white capital income, determines the r_I which, in turn for a well-behaved neoclassical production function, determines the capital-labor ratio $(K/L)_I$, and hence L_I (since $\overline{K}_I$ is assumed given). The second constraint then forces a level for c. And the final constraint fixes w^E. For any given return to industrial capital, the white wage rate (and hence white labor income) is determined.

This income possibility curve is illustrated in Figure 1, with white labor income on the vertical axis and white capital income on the horizontal axis. Different slopes and curvatures are possible.[21] Only two things

[16]Consideration of a model with two kinds of labor greatly alters this discussion of efficiency The wage rate difference ($w^E > w^B$) then does not imply that identical labor is being differentially rewarded, w^E is the wage rate of *skilled* labor and w^B of *unskilled* labor—the extremely biased process by which only whites become skilled may be appallingly inequitable, but it is no longer necessarily inefficient. But other new sources of inefficiency enter, here it must suffice to note them: i) If the ability to absorb education (defined as you will) of the *most* apt excluded black is greater than that of the *least* apt included white, there appears inefficiency in the sense that any *given* quantity of training is not achieved at least cost; and ii) If the skill differential (i.e., $w^E - w^B$) is set too high, employers will employ too few skilled workers—in the sense that the opportunities for training labor whose additional productivity exceeds the marginal training cost are not exhausted.

[17]At least, as an *explicit* transfer policy. Taxation of capital to expand public employment of whites (and coloreds) is partly so motivated, but, unless the public employment is productive, this introduces another source of inefficiency.

[18]From a 1924 government statement, quoted in Muriel Horrell, p. 57.

[19]In South Africa, the tradeoff is also a cultural and political problem. The government and white labor are predominately "Afrikaner" (i.e., of Dutch descent), and the capitalists "English." That apartheid policies are more fervently backed by Afrikaners is not inconsistent with economic advantage.

[20]L_I refers to the total industrial labor force, $L_I^B + L_I^E$.

[21]Note the end points of the income possibility curve. At the northwest, it must cease once r_I has fallen so low, and hence $(K/L)_I$ risen so high, that c must equal one to achieve full white employment, given $\overline{K}_I$. At the southeast, there is no practical interest in considering $w^E < w^B$. As we move from the point where $c = 1$, by lowering w^E and hence raising L_I, the income of capitalists must rise; whether the income of white labor rises or falls depends on the magnitudes of second derivatives of $X_I(\cdot)$.

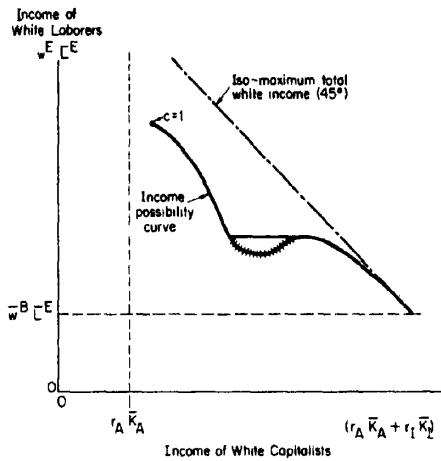



FIGURE 1

are certain: i) that the entire curve falls inside (i.e., below) the dot-dashed 45° line which shows the maximum attainable total white income (achievable only if $w^E = \bar{w}^B$); and ii) that if there are upward-sloped segments, they are dominated (as the hatching of Figure 1 shows).

Thus there are conflicts between capital and labor in this white dominated economy, even when the single goal is so seemingly straightforward as maximization of white incomes. But such maximization is not the only goal; and the existence of other goals introduces further sources of conflict. It is difficult to distill the essence of "the" goals of South African whites: their policy as well as their philosophy displays an inconsistency that is inevitable where the black presence is deplored while the white living standard depends upon it. Nevertheless, two broad kinds of economic goals seem to emerge.

One, the concept of *apartheid* has economic as well as political and social meaning. It means that black labor should work separately from white labor. In part, this simply means that instead of increasing the fraction of black workers in all industry, new factories with a high percentage of black workers should be located away from the white cities (i.e., the "border areas industrialization" program; see Trevor Bell). But it also has meant a continued resistance to the rising importance of black labor in agriculture and, even more, industry. This resistance stems both from labor's fear that white full employment is threatened and from a more profound fear of excessive white economic "dependence" on black labor which could eventually endanger the whites' political and social dominance. Thus, one goal of the Southern African-type economy is a reduction in the level (or growth rate) of black employment outside of the reserves.

And second, South Africa has long encouraged the growth of industry at an even faster rate than the natural forces of economic development evoke. In this, it is no different from nearly every developing country of the past two centuries; and it has employed the standard policies of tariff protection, tax advantage, and direct subsidy to encourage the industrialization. This goal stems partly from the usual beliefs about the inferiority of primary products and the positive externalities and dynamic benefits generated by industry, but in South Africa there is much more. Policies to encourage industry emerged at the same time as excessive fragmentation of farm ownership was creating a class of "poor whites" in the cities. As whites refused to do the unskilled rural work (reflected in the model by the absence of L_A^E), it was necessary to encourage a rapid growth of demand for white labor in industry to insure a politically feasible distribution of the rising average white standard of living.

Finally, since World War II, changes in international attitudes and the political structure of Africa have generated a fear of isolation through economic sanctions; industrialization reduces the dependence of South Africa on its mineral exports and its industrial imports. Although trade is ignored in our model, these concerns can all be reflected in the model as a goal of higher levels (or higher growth rates) for industrial output, industrial capital, and the rate of return to capital in industry.

These various goals are summarized, in simplistic fashion, in Table 1. In the column labeled Goal, the sign indicates the direction of change desired, *ceteris paribus*, in various relevant variables of the model (and

TABLE 1—RELATION OF POLICY INSTRUMENTS TO GOALS

Variable	Goal	Parameter Change			
		$\Delta w^E > 0$	$\Delta c > 0$	$\Delta w^B > 0$	$\Delta p > 0$
X_I/X_A	+	–	–	?	+
L_I^E	+	–	?	?	+
L_I^B	–	–	–	?	+
L_A^B	–	+	+	?	–
$L_I^B + L_A^B$	–	?	?	?	?
K_I/K_A	+	–	–	?	+
r_I	+	–	–	–	+
r_A	+	0	0	–	0
X_R	+	?	?	?	?
L_R	+	?	?	?	?

a few combinations of variables, namely, X_I/X_A, $L_I^B + L_A^B$, and K_I/K_A). The impact on these goals of changes (Δ) in each of the three parameters (w^E, c, and w^B) are shown in the next three columns (the final column, labeled $\Delta p > 0$, is discussed shortly); positive parameter changes are examined because these are the preferred directions of change.[22] The signs of Table 1 are readily derived since the analysis follows closely that of Section II. The conflicts are clear. For many desirable parameter changes, the impact on the goals is uncertain without detailed quantitative information about the economy's technological and behavioral relations. And, where the qualitative effect is certain, it is usually in the wrong direction from the viewpoint of the goals.

Does price policy offer an escape from these uncertainties and conflicts? Assume that agricultural output (i.e., both X_A and X_R) is not only the *numeraire* good but also the output in which both wage rates (i.e., w^E and w^B) are denominated, and that the government's tariff policy permits it to vary the domestic price of industrial output. Explicit consideration of the internal industrial price requires change in the model only of equation (3) to

$$X_I = pX_I(K_I, L_I^B + L_I^E) \tag{3'}$$

where p is the price of industrial output. An increase in p shifts the domestic value of production function upward; from the viewpoint of producers, such a rise is equivalent to a Hicks-neutral technological improvement. Thus, an increase in p will raise the rate of return to capital and reduce the capital-labor ratio in industry. Since the rise in p has no impact on the rate of return to capital in the agricultural sector, capital is drawn into industry. This together with the lower capital-labor ratio insures an increase in industrial employment (of both black and white labor, since c is assumed unchanged). These changes are summarized in the final column of Table 1. Again there is uncertainty about the net impact on black labor in agriculture *and* industry. But with this exception, the policy of raising p contributes quite satisfactorily to the various goals of white policy.

While it is fairly realistic to treat black wages as being fixed in terms of food, the rise in p surely lowers real white wages; moreover, if white full employment has already been achieved, the rise in p creates excess demand for whites. However, it is possible to combine rises in p and w^E so as to maintain a constant level of white employment. This result requires that the proportionate increase in the capital stock of industry be exactly equal to the proportionate rise in the capital-labor ratio there.[23]

[22] That higher w^E and c are preferred is obvious. The case for higher w^B is less clear; it rests on the whites' hopes for labor stability, international respect, and urban quiet.

[23] Since $L_I = \bar{L}^E/c = K_I/(K/L)_I$.

At the new equilibrium with higher p and w^E but L_I unchanged: i) the rate of return to capital in industry is higher; ii) output per worker in industry is higher; and iii) total industrial output (at world prices as well as at domestic prices) is higher. Note also that total black employment off the reserves is reduced, since their industrial employment is unchanged and their agricultural employment is reduced as agricultural capital declines and no change in $(K/L)_A$ occurs.

From the viewpoint of the white voters, this combined policy, rising p and rising w^E, would seem an almost ideal solution. However, one must look not at the *nominal* but at the *real* income changes involved. Assume that all whites, both laborers and capitalists, buy some food (at unchanged prices) and some manufactures (at now higher prices). Clearly, agricultural capitalists lose since not even their nominal rate of return has risen. The real rate of return to capital in industry may also have fallen; the nominal rate rises less than industrial prices, so if these capitalists spend most of their income on manufactures, they will be worse off. And the nominal wage rate of white workers has risen by more than industrial prices, so no matter what their consumption pattern, the real white wage rate has risen.[24] Thus, the simultaneous rise of p and w^E does increase industrial output and reduce white dependence on black labor, but it does so at the cost of possibly serious income redistribution among whites, from capital (especially agricultural capital) to labor.

Moreover, there is a loss in the total output of the economy, measured in world prices—that is, the sum of X_R, X_A, and X_I is reduced. The movement of labor from agriculture to the reserves cannot increase output since the marginal product of each black worker in agriculture w^B must have been at least as high as his opportunity cost in the reserves b, after adjustment for any nonpecuniary differences. The movement of the *first* unit of capital from agriculture to industry involves no loss since the owner must have been indifferent between his earnings in agriculture and in industry. As subsequent capital flows occur, there is no change in the rate of return to capital in agriculture, since capital and labor are withdrawn together there (at constant $(K/L)_A$). But the addition of this capital to industry is made with a constant industrial labor force, and hence the rate of return to capital in industry must fall. In short, there is a decline in total output in world prices, that is, the *real* output. The gainers (white labor) gain less than the losers (white capital) lose.[25]

This section can be summarized in a sentence. The complete white domination of the economy and its policy parameters does not free whites from awkward conflicts and contradictions between the subclasses of white labor and white capital and between the different policy goals which whites simultaneously seek.

IV. Dynamics

As a first step to uncovering the growth paths of the Southern African-type economy, let us ignore (quite unrealistically) technical change and assume (realistically) that none of the fruits of accumulation are passed on to black workers.

Consider first the path of balanced growth, by which I mean that the capital stock, employment, and output in both agriculture and industry all grow at the same constant rate. If the black wage rate is held constant over time and unlimited sup-

[24] When the value of production function increases by a factor p, an equal proportional increase of the weighted-average industrial wage $cw^E + (1 - c)w^B$, would imply that the new tangency be at the same capital-labor ratio as before Since the capital-labor ratio rises, the proportionate increase in the weighted-average wage rate exceeds the proportionate increase of industrial prices But the white wage rate is only a part of that weighted average, and the rest (i.e., $(1 - c)w^B$) does not rise at all. So w^E rises, a fortiori, by proportionately more than industrial prices.

[25] This gives a somewhat inaccurate picture of South African tariff policy and problems. Actual policy has protected *both* industry *and* agriculture while taxing the exports of the mining sector. Our model is not large enough to analyze this situation, but it does point out the potential conflict evoked by tariff policy.

plies of black labor continue to be available from the reserves, growth in the agricultural sector occurs at a constant rate of return to capital. Thus, balanced growth (of labor, capital, and output) in agriculture can occur at any growth rate.

If the behavior of the capital market in allocating new capital between agriculture and industry is unchanging over time, balanced growth of the capital stocks K_A and K_I requires that the relative rates of return to capital r_A/r_I remain constant. Since r_A is constant for any agricultural growth rate, r_I must also remain constant if balanced growth is to occur. But this requires that the weighted-average industrial wage rate $cw^E + (1 - c)\overline{w}^B$ remains constant over time. Since $\overline{w}^B$ is constant, this means that w^E can rise only if c falls.

What happens to c depends on the relative rates of growth of the total capital stock in the economy and the white labor force. Balanced growth means that the rate of growth of the total capital stock is also the sectoral rate of growth of capital and employment in each of agriculture and industry. But the rate of growth of white employment in industry is given at the exogenous growth rate of the white labor force. Thus, when the total capital stock grows at a more rapid rate than the white labor force, c must decline.[26] And, since the weighted-average wage rate in industry must remain constant, the white wage rate must rise.

Balanced growth would seem a heart-warming proposition for whites. The rate of return to their capital is not falling and the wage rate of their labor is rising. The conflict arises with respect to the other goals of Southern African development—that is, a reduced dependence on black labor and an increased industrial share of total output. Balanced growth does not support the latter goal by definition; and even balanced growth requires the growth of black employment in industry at a faster rate than capital and output there.

The introduction of technological change alters the balanced growth analysis very little. With $\overline{w}^B$ still constant, technical progress in agriculture, assumed disembodied for simplicity, raises the rate of return to capital there. Balanced growth requires that the rate of return to capital rise equally rapidly in industry. Depending on whether technical progress is greater in industry or agriculture, the weighted-average industrial wage rate will have to rise or fall. In this case, blacks are being denied any share in the fruits of either the capital accumulation or the technical progress, but it is not clear who the white beneficiaries are. Capitalists surely gain, since rates of return to capital rise; but whether white laborers share in the gains due to such progress depends on i) the relative rates of technical advance in industry and agriculture, and ii) the degree of factor substitutability in the industrial production function. The income gains due to technical progress are not automatically divided between white capital and white labor in politically acceptable shares.

Finally, note the impact on this economy of inflows of foreign capital or of increased white immigration. This is easily done, since the former means basically a more rapid accumulation of capital and the latter a more rapid rate of growth of the white labor force. The impact of each on the paths of output, employment, etc. is easily derived, but the interesting question is how these changes affect the well-being of the blacks. The answer is: little. The black wage rate w^B can be set, within limits, wherever the white policymakers wish. Higher rates of capital accumulation or lower rates of white labor force growth will, *ceteris paribus*, reduce c, and the former will raise L_I as well. If one accounts participation of blacks in the "modern" (i.e., nonreserve) sectors as adding to their well-being, then there is some positive impact.[27]

[26] Over 1911–70, real South African Gross Domestic Product (*GDP*) grew at over 4 percent per year, while the white population grew at less than 2 percent.

[27] In reality, more rapid capital accumulation in industry means not only a more rapid decline in c but also increased training and better access to skilled jobs for blacks. The model, with but one kind of labor, cannot treat this. The model also cannot consider the possibility that investment in black human capital or by foreign capital will gradually force a rise in the average black wage rate.

The desire in South Africa for faster rates of foreign capital inflow and for white immigration points up anew the dynamic conflicts of goals and means there. Greater capital inflow is sought in order to accelerate industrialization, even though it means a more rapid decline of c. And greater white immigration is sought to repair the damage to c, even though it in turn reduces the rate at which w^E can rise.

V. Exploitation

Finally, the model yields insight into what is ultimately the most interesting question: exactly how does Southern African exploitation of blacks by whites occur? In order to answer this question, however, we must carefully define the word "exploitation." One straightforward and statistically sensible measure is the income disparity. By most measures, South African income is the most inequitably distributed in the world.[28] A second and analytically more interesting concept is the divergence of factor prices and marginal products. By this approach, the exploitation is seen to arise in the industrial sector, where each white laborer whose wage rate exceeds his marginal product by $(1 - c)(w^E - w^B)$ is seen to exploit each black laborer, whose (identical) marginal product exceeds his wage rate by $c(w^E - w^B)$. A third approach to exploitation is the comparison of the political solution actually reached with the efficient competitive solution discussed earlier. Then the white workers are seen to be exploiting not blacks (who cannot be much exploited if w^B is near b which in turn is near subsistence) but white capitalists.

There is also a fourth, and I think more interesting, way of looking at the question of exploitation, namely, that it arises from the whites' ability to integrate the black and white economies to the whites' advantage. Then the allocation and distribution solution of the Southern African-type economy should be compared with an efficient competitive system in which white capitalists (and laborers) do not have the opportunity of using black labor, that is, a complete economic apartheid of black and white factors of production. The manner of exploitation is most sharply seen in the industrial sector. To simplify the analysis at first, it will be assumed that capital is completely immobile between sectors; then, the marginal product of industrial labor can be plotted against labor, as in Figure 2. Consider first the situation where no black labor can be used. Full employment of the white labor force $\overline{L}^E$ requires a wage rate of w_s^E (s refers to segregation). The total earnings of the (white) industrial capital are represented by the area α, and the total earnings of the white industrial labor by the sum of the areas $\beta + \gamma + \delta$. Now consider the situation where black labor can be hired at a wage rate of w^B. Suppose L_I^i (i for integration) total laborers are hired. The job-reservation ratio c_i is determined by the need to maintain white full employment (i.e., $c_i = \overline{L}^E/L_I^i$); and the white wage rate w_i^E is determined by the need for the weighted-average wage rate to equal the marginal product of labor at L_I^i. Now, the total earnings of industrial capital are represented by the areas, $\alpha + \beta + \epsilon$; integration clearly increases the earnings of capital. The total industrial labor earnings

FIGURE 2

[28]The poorest 40 percent of the population receives only 6.2 percent of the income, the lowest percentage of the sixty-six countries surveyed by Montek Ahluwalia, pp. 8–9.

are represented by the areas, $\gamma + \zeta + \delta + \eta$; since the black workers (i.e., $L_I^I - \overline{L}^E$) must be paid a wage rate of w^B, the white labor earnings are represented by the area, $\gamma + \zeta + \delta$. It is not certain whether integration has benefited white workers—this depends upon whether $\zeta \gtreqless \beta$. But it is certain that *all* whites together, laborers plus capitalists, gain by integration to the extent of the areas $\epsilon + \zeta$.

We can now see the source and extent of the exploitation. When the races are separate, the white economy is relatively capital abundant, as indicated in Figure 2 by the low earnings of capital α, and the high wage rate w_I^E. The black economy has no capital and the marginal product of black labor is low. The chance to integrate these very different economies offers great potential for gain. The model shows that exploitation occurs in the sense that this gain is captured by whites.[29]

The exploitation in agriculture is more elemental. There without the merger, there is no white labor at all, so the entire income of white capital derives from its capture of the gains from "trade" with black labor. Here, too, unless w^B exceeds b, white capital extracts *all* the gains, raising its income share from zero. Of course, if racial segregation were established, eventually either white labor would move to agriculture or agricultural capital would move to industry. But the qualitative results are not changed and the quantitative magnitudes are exacerbated. Regardless of whether white labor or agricultural capital moves, the relative capital abundance of white industry is increased, so that the potential gains from integration would be even greater.

It should be noted in closing that this discussion of exploitation is entirely static and ahistorical. If the "black economy" had not been "merged" with the white economy, it might have, but presumably would not have, gone on forever as a capital-less, technically stagnant reserve.[30] Sensible economic policy by independent blacks could have eventually created saving, growth, and diversity in the black economy. The merger of the black economy into the white economy and its consequent subjugation to white policy have denied this possibility; white development efforts have gone preponderantly to the augmentation of technological capacity and white human and physical capital in the modern sectors where whites could gain most from access to cheap, unskilled black labor. Meanwhile, the continued low productivity of the reserves has made a continued low wage rate for black labor possible. This brings out the ultimate paradox of white policy: while the demands of internal politics evoke a rhetoric of "separate development," the continued exploitation of blacks requires both their integration and their non-development.

[29] Recall that the actual allocation is inefficient, so that some of the potential gains from integration go unrealized (namely, the area θ in Figure 2).

[30] Similarly, without blacks to exploit, the white economy might nevertheless have become an Australia (or an Argentina)

REFERENCES

M. S. Ahluwalia, "Income Inequality: Some Dimensions of the Problem," in Hollis B. Chenery et al., eds., *Redistribution with Growth*, London 1974, ch. 1.

Trevor Bell, *Industrial Decentralisation in South Africa*, Capetown 1973.

G. V. Doxey, *The Industrial Colour Bar in South Africa*, Capetown 1961.

W. Elkan, "Migrant Labor in Africa: An Economist's Approach," *Amer. Econ. Rev. Proc.*, May 1959, *49*, 188–97.

S. Enke, "South African Growth: A Macro-Economic Analysis," *S. African J. Econ.*, Mar. 1962, *30*, 34–43.

P. R. Fallon and P. R. G. Layard, "Capital-Skill Complementarity, Income Distribution, and Output Accounting," *J. Polit. Econ.*, Apr. 1975, *83*, 279–302.

Sally H. Frankel, *The Economic Impact on Under-Developed Societies*, Cambridge 1959.

Muriel Horrell, *Legislation and Race Relations*, Johannesburg 1971.

J. B. Knight, "A Theory of Income Distribution in South Africa," *Bull. Oxford Univ. Instit. Econ. Statist.*, Nov. 1964, *27*, 289–310.

——— **and M. D. McGrath,** "An Analysis of Racial Wage Discrimination in South Africa," *Bull. Oxford Univ. Instit. Econ. Statist.*, Nov. 1977, *39*, 245–71.

D. Yudelman, "Industrialization, Race Relations and Change in South Africa: An Ideological and Academic Debate," *African Aff.*, Jan. 1975, *74*, 82–96.

Optimal Tax Schedules and Rates: Mirrlees and Ramsey

By ROBERT COOTER*

The formulation of a utilitarian tax philosophy in the nineteenth century by John Stuart Mill and others did not lead to the conclusion that there should be a progressive income tax.[1] That conclusion was reached by Francis Edgeworth when he demonstrated that equal marginal sacrifice requires equal after-tax income for different individuals in the absence of "announcement effects" from taxation.[2] The proof established a presumption that a utilitarian tax philosophy requires progressive income taxation, which lies behind contemporary ability-to-pay arguments for progressivity.

Edgeworth recognized the important effects of income taxation upon work effort, but he was unable to incorporate these effects explicitly into his mathematics. Recent formulations have corrected this omission and a revision of utilitarian tax philosophy is underway which is as fundamental as what took place after Edgeworth first obtained his results. The pioneer in this work is James Mirrlees.[3] Unfortunately, much of the theoretical literature is technical and abtruse. This paper derives the fundamental theorems by methods which are intelligible to anyone familiar with the Maximum Principle.[4] There are also modest extensions of known results. The first part of the paper deals with the simple two-good model and takes advantage of many simplifying assumptions; the second part concerns optimal tax schedules for several commodities; the third part derives Ramsey-type rules for the taxation of commodities which are susceptible to tax rates but not schedules, in a context where there are many persons and many goods.

I. Optimal Income Taxation with Two Goods

The original papers on optimal income taxation concerned the effect of taxation upon work effort or investment in education. My exposition will be based upon the effort-incentive model, in which the individual responds to the tax by adjusting his labor supply. The individual's utility function is written

$$u = u(-y, x)$$

where x is consumption, y is labor, and u is assumed to be concave and twice differentiable. Individuals differ with respect to their productive skill n, which is assumed to be equal to their gross wage rate. Total wages before taxation for an "n-person" are the product of his wage and labor supply: $z = yn$. Consequently, we may write the problem of individual choice by an n-person

$$\max_{z,x} u(-z/n, x)$$

subject to $z - T(z) - qx \geq 0$

where $T(\cdot)$ is the income tax schedule; q is the price of x, which may as well be unity in the two-good model. I shall assume an interior solution, so the first-order conditions are

$$(1) \qquad \begin{aligned} u_1 &= n(1 - T'(\cdot))\Psi \\ u_2 &= \Psi \\ z - T(z) - x &= 0 \end{aligned}$$

*Department of economics, University of California-Berkeley, and The Institute for Advanced Study. The study was funded in part by a grant from the National Science Foundation.

[1]"Mill favored neither an income tax nor (except in special circumstances) the application of progression" (Harold M. Groves, p. 33)

[2]See Harold M. Groves, ch. 6; also see Edgeworth. I assume that the necessary tax revenues are not realized before taxation has equalized incomes.

[3]The original papers were by Mirrlees (1971) and Fair (1971); a major extension is found in Mirrlees (1976).

[4]There are many expositions of the Maximum Principle. For example, see Robert Dorfman or Kenneth Arrow and Mordecai Kurz

Note that Ψ is the Lagrange multiplier which indicates the marginal utility of expenditure. Subscripts on functions indicate partial derivatives.

A feature of this model is that an individual's productive skill or hourly wage is assumed to be independent of the tax structure;[5] consequently, there is a distribution $f(n)$ which is unaffected by the control variable. I also assume $f(n)$ to be continuous. Let x_n and y_n indicate the consumption and labor of an n-person. The subscripts serve to remind us that the variables depend upon the person's productive skill; I omit these subscripts in the text to avoid cluttered notation. The government's problem is to maximize an additive social welfare function $G(u(\cdot))$ subject to the government's revenue need $\bar{R}$ and the labor-supply response of individuals:

$$(2)\qquad \max_{z,x,T} \int_{N_1}^{N_2} G(u(-z/n,x))f(n)\,dn$$

subject to

$$(3)\qquad \bar{R} = \int_{N_1}^{N_2} T(z)f(n)\,dn$$

$$(4)\qquad T(z) = z - x$$

For convenience we take N_1 and N_2 to be nonnegative and finite.[6] The government chooses the tax/subsidy schedule $T(\cdot)$; it also chooses the consumption schedule x_n and the income (hence labor supply) schedule z_n, but choice is constrained by the individual maximizing conditions (1).

We may use the Maximum Principle to solve this problem if we reformulate the constraints as differential equations. First we combine (3) and (4); and then differentiate the combined equation so that the government's budget constraint can be written as a differential equation and a terminal condition:

$$(5)\qquad DR = (z - x)f(n)$$

$$(6)\qquad R_{N_2} = \bar{R}$$

(The symbol D is used throughout for the derivative d/dn.) I substitute (5) and (6) for (3) and (4) in the maximization problem; by this step I eliminate one control variable, namely $T(\cdot)$, and add a state variable, namely R_n. Of course $T(\cdot)$ is still chosen implicitly since it is the difference between income and consumption for each person.

The next step is to write the individual maximizing conditions (1) as a differential equation. Define the utility of an n-person when z and x are optimally chosen to satisfy (1):

$$(7)\qquad v_n = \max u(-z/n,x)$$

By differentiating with respect to n we obtain

$$(8)\quad Dv = (-u_1 Dz/n + u_2 Dx) + u_1 z/n^2$$

I wish to simplify (8). Differentiate the individual's budget constraint with respect to n and substitute the other two conditions from (1) into it, which leaves the equation

$$(9)\qquad -u_1 Dz/n + u_2 Dx = 0$$

In view of this fact, (8) may be simplified.

$$(10)\qquad Dv = u_1 z/n^2$$

Equation (1) implies (10); it will be assumed that the converse is also true. The special circumstances under which this equivalence does not hold are a technical detail which is relegated to a footnote.[7] By this assumption we may substitute the constraints (7) and (10) for (1). In fact we may simplify further by inverting (7) $z = g(v,x,n)$. (v_n is monotonically decreasing in z_n, so inversion is permitted.) This inversion enables us to eliminate z_n as a control and use v_n as a state variable. The final form of the problem is

$$\max_{x,v,R} \int_{N_1}^{N_2} G(v)f(n)\,dn$$

subject to

[5]Relaxing this assumption makes little difference to simulation results with linear tax schedules. See Martin Feldstein.

[6]In fact no one has an hourly wage which is negative or infinite.

[7]The individual maximizing conditions (1) are sufficient for (10); the problem is that (1) is not necessary for (10). The problem cases occur along a path z_n, x_n which satisfies (10) but not (1). Such a difficulty would arise along a path of constant consumption and income: $0 = Dx = Dz$. This problem is discussed in Mirrlees (1976).

$$Dv = u_1(-g(v,x,n)/n,x)\cdot g(v,x,n)/n^2$$
$$DR = (g(v,x,n)-x)f(n)$$

with the terminal condition $R_{N_2} = \bar{R}$ and the initial condition $R_{N_1} = 0$, where x_n is the control and the state variables are v_n and R_n; by straightforward application of the Maximum Principle we obtain the Hamiltonian and first-order conditions:

$$H = G(v)f(n) + \mu u_1(\cdot)g(\cdot)/n^2 + \lambda(g(\cdot) - x)f(n)$$

I: $\partial H/\partial x_n = 0$

II: $\partial H/\partial v_n = -D\mu$

III: $\partial H/\partial R_n = -D\lambda$

IV: $Dv_n = u_1\cdot g(\cdot)/n^2$

V: $DR_n = (g(\cdot)-x)f(n)$

VI: $\mu_{N_1} = \mu_{N_2} = 0$ (transversality conditions)

VII: $R_{N_1} = 0; R_{N_2} = \bar{R}$

All of the theorems are obtained by interpretation of the necessary conditions I–VII.

THEOREM 1: *A person with higher productive skill enjoys at least as high utility as a person with lower productive skill; formally,* $Dv_n \geq 0$.

PROOF:

The theorem follows directly by interpreting the signs in condition IV. This first theorem reflects the fact that a person with higher ability can always earn the same income and pay the same taxes as someone with lower ability, but enjoy greater leisure. Theorem 1 is the motivation for Figure 1 in which equal utilities is only achieved at the origin by tax rates which are so high that neither person works. Figure 1 illustrates a limitation of income taxation, namely that the point of equal utility for everyone is off the Pareto frontier.

THEOREM 2: *The marginal tax rate on income is less than one;* $T' < 1$.

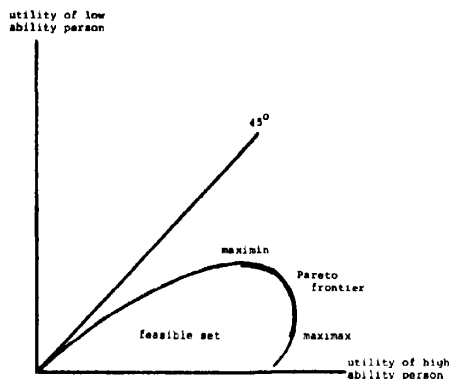



FIGURE 1. PARETO FRONTIER

PROOF:

Condition IV is equation (10), which implies (1) by assumption. The first condition in (1) can be written

$$T' = 1 - u_1/\Psi n$$

In an income interval where $T' > 1$, an increase in gross income produced by a sacrifice of leisure would result in a decrease in income net of tax. No one would exchange labor for a decrease in income, so no one's preferred z will fall inside this interval. Either everyone will choose z below that interval, in which case the marginal tax in the interval may as well be reduced, or else someone's utility function will have two local maxima, in which case there is a nonconvexity. The theorem tells us more about the limiting assumptions of the model than about reality.

THEOREM 3: *The marginal tax rate is nil for the person of lowest ability and for the person of highest ability;*[8] $T'(z_{N_1}) = T'(z_{N_2}) = 0$.

PROOF: (i) Expand condition I to obtain

[8]The original observation that marginal taxes are nil at the upper end was made by Efraim Sadka and Edmund Phelps; the observation that marginal taxes are nil at the lower end is attributed to Seade (1975).

$$0 = \frac{\partial H}{\partial x} = \mu\left\{\frac{\partial}{\partial x}(u_1 g/n^2)\right\} + \lambda(g_2 - 1)f$$

(ii) Observe that

$$g_2 = \frac{dz}{dx} = \text{marginal rate of substitution}$$
$$= +1/(1 - T') \text{ from the individual maximizing conditions (1)}$$

(iii) Combine (i) and (ii) to obtain the marginal income tax rate:

$$T' = \frac{-\mu\{\cdot\}}{\lambda f g_2}$$

(iv) The theorem follows immediately from (iii) and the transversality conditions VI

$$\mu_{N_1} = \mu_{N_2} = 0$$

The transversality conditions require the shadow price μ_n to be nil at the upper and lower ends of the integral because the state variable v_n is unconstrained by initial or terminal conditions. (At the optimum the vector of shadow prices must be orthogonal to the tangent plane on the manifold representing the initial or terminal constraints.)

There is a simple interpretation for the conclusion that marginal tax rates should be nil for the person of greatest productive skill. Consider any tax schedule with $T'(z_{N_2}) > 0$; now construct another tax schedule identical to the first for all $z \leq z_{N_2}$ and set the right derivative $T'(z_{N_2}) = 0$, as shown in Figure 2. Tax collections from every individual are the same under both schedules, so no one of lower ability is worse off. However, the person of highest ability has a larger opportunity set under the second schedule and he may choose to work more; since his welfare is improved, the original schedule cannot have been optimal under the Pareto criterion or any social welfare function which gives positive weight to his interests.

Proving that the tax schedule is nondecreasing is difficult in spite of the triviality of the result. The following theorem is

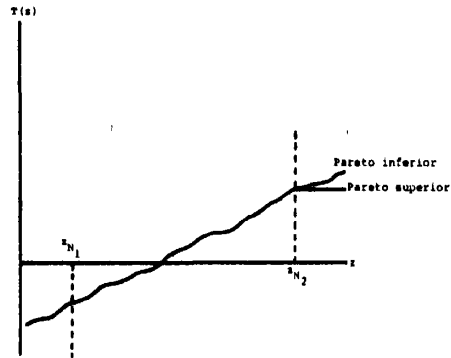



FIGURE 2 ILLUSTRATING THEOREM 3

proved in the Appendix:

THEOREM 4: *The optimal income tax schedule is nondecreasing* ($T' \geq 0$) *if the following conditions are met:*
(i) *Cross partial is nonnegation:*[9] $u_{12} \geq 0$
(ii) *The marginal social value of leisure decreases with ability:* $d(G'u_1)/dn \leq 0$

If one person enjoys a higher hourly wage than another, then he enjoys higher utility according to Theorem 1. An implication of assumption (i) is higher consumption increases the value of leisure; assumption (ii) requires the tax system to assign less weight to his leisure. In other words, assumption (ii) requires the tax system to distribute leisure in a way which offsets unequal consumption, which is why the assumption is appealing.

From Theorems 3 and 4 we can conclude that there is a zone of increasing marginal tax rates and a zone of decreasing marginal rates; they do not increase everywhere as might be supposed from the Edgeworth-Pigou tradition. Furthermore, the zone of decreasing marginal rates occurs at a higher income level than the zone of increasing rates. If the marginal tax rate for the optimal schedule has a single maximum, then the schedule of marginal rates is concave,

[9]The role of the cross partial's sign in determining the nonnegativity of the marginal tax rate is discussed by Sadka.

with an interior maximum. Of course the theorems tell us nothing about the rate at which the optimal marginal rates approach zero; simulations must determine whether Theorem 3 is useful or a curiosity. What conclusions can be drawn about progressivity? Progressivity means that an increasing proportion of income is paid in taxes; we can conclude nothing about progressivity from the theorems because they refer to marginal rates only and say nothing about the intercept (tax on zero income). It seems reasonable from simulations (and can be proven under particular circumstances) that the tax rate on zero income will be negative; a demogrant will be paid to everyone.[10] The theorems say nothing about its size so I can say nothing about average tax rates.

The two-commodity model captures the disincentive effects of income taxation upon work effort. An obvious criticism of this model is that it tells nothing about a general system of taxation; I shall remedy that in the next section.

II. Tax Schedules for Many Goods

It is administratively possible to have a tax schedule for goods other than income. For example housing is subsidized (negative tax) for some people at a rate which varies with their expenditure upon it. In this section x is interpreted as a vector of goods, each of which is susceptible to a tax schedule. However, it is unrealistic to assume that a tax schedule is administratively feasible for every commodity; for most commodities government must be content with tax rates, particularly when resale is possible. Let $c = (c_1, c_2, \ldots, c_c)$ be a vector of commodities which are susceptible to tax rates but not tax schedules. Mirrlees' problem is to optimize the tax schedules for z and x, and Ramsey's problem is to optimize the tax rates for c. I set up the control problem and solve Mirrlees' problem in this section; in the next section Ramsey's problem will be solved.

[10]This is a standard result when the optimal tax is linear. See Eytan Sheshinski or the author and Elhanan Helpman.

I wish to maximize the weighted sum of utilities subject to the government's budget constraint and the conditions for individual utility maximization. As in Section I, the key to the problem is formulating the constraints as differential equations. First consider the problem of individual utility maximization. In the general formulation an n-person must solve

$$\max_{z,x,c} u(-z/n, x, c) \tag{11}$$

subject to

$$-z + T(z) + \sum_i (p_i x_i + Q^i(x_i)) + \sum_j q_j c_j = 0 \equiv m$$

where
z = labor income
T = income tax
$p = (p_1, p_2, \ldots)$, seller's prices
$x = (x_1, x_2, \ldots)$, a vector of consumption goods
Q^i = tax schedule for x_i
$q = (q_1, q_2, \ldots)$, buyer's prices
$c = (c_1, c_2, \ldots)$, a vector of consumption goods
m = net expenditure

Equation (11) can be expressed as a differential equation if we follow the same tack as the two-good case. Define utility as a function of the optimally chosen commodity bundle (satisfies (11)):

$$v_n = \max u(-z/n, x, c) \tag{12}$$

Differentiate this function and the budget constraint in (11) with respect to n; combine these results with the first-order conditions in (11) to obtain our familiar differential equation:

$$Dv = \frac{\partial}{\partial n} u(-z/n, x, c) = u_1 z/n^2 \tag{13}$$

Notice that (11) implies (13); it will be assumed that the converse is also true. (This assumption was discussed for the two-good case in fn. 7.)

It is customary to set up Ramsey's problem by using the indirect utility function rather than the direct utility function, because the government chooses consumer prices rather than quantities. Similarly, we shall use the expenditure function so that we can make the vector of buyer's prices q into the control variable, rather than the quantities c. Consider any system of tax schedules T and Q. What is the minimum expenditure which realizes a particular utility level v, given z/n, x, and q?

$$(14)\quad \min_c -z + T(z) + \sum_i [p_i x_i + Q^i(x_i)] + \sum_j q_j c_j$$

subject to $u(-z/n, x, c) \geq v$

Let the solution be the expenditure function

$$(15)\qquad m = m(q, v, -z/n, x)$$

with the partial derivatives being Hicksian demand

$$(16)\qquad m_j(\cdot) = c_j$$

By substitution into (13) we can eliminate c from the differential equation,

$$(17)\qquad Dv = (z/n^2)u_1(-z/n, x, m_q)$$

The differential equation we have been seeking is (17).

We can write the revenue constraint as a differential equation and a terminal condition just as in the two-good case. The final formulation of our general problem is written

$$\max_{z,x,v,R,q} \int_{N_1}^{N_2} G(v)f(n)\,dn$$

subject to

$$Dv = (z/n^2)u_1(-z/n, x, m_q(\cdot))$$

$$DR = \left(z - \sum_i p_i x_i - \sum_j p_j m_j(\cdot)\right)f$$

and the conditions

$$R_{N_1} = 0$$
$$R_{N_2} = \overline{R}$$

The controls are z, x, and q; the state variables are v and R.

I find the following first-order conditions by application of the Maximum Principle:

$$H = G(v)f(n) + \mu\frac{\partial u}{\partial n}(\cdot) + \lambda(z - \sum_i p_i x_i - \sum_j p_j m_j(\cdot))f$$

$$\text{I: } 0 = \frac{\partial H}{\partial z} = \frac{\partial H}{\partial x}$$

$$\text{II: } -D\mu = \frac{\partial H}{\partial v}$$

$$\text{III: } -D\lambda = \frac{\partial H}{\partial R}$$

$$\text{IV: } Dv = \frac{\partial u}{\partial n}$$

$$\text{V: } DR = \left(z - \sum_i p_i x_i - \sum_j p_j m_j\right) f$$

$$\text{VI: } \mu(N_1) = \mu(N_2) = 0$$

$$\text{VII: } R_{N_1} = 0; R_{N_2} = \overline{R}$$

Different values may be chosen for z, x, v, and R at different values of n; the vector of prices q may also be chosen, but q is invariant with respect to n. The first-order condition on the choice of the q's is obtained by the usual reasoning in the calculus.

$$\text{VIII: } 0 = \frac{\partial J}{\partial q_k} \text{ all } k$$

where $J = \int_{N_1}^{N_2} \{G(v)f(n) + \mu(n)\left(-Dv + \frac{\partial u}{\partial n}\right) + \lambda\left(-DR + (z - \sum_i p_i x_i - \sum_i p_j m_j)\right)f(n)\}dn$

Theorems 1 and 2 and their proofs carry over from the two-good case without any change, so there is no need to repeat them. Theorem 3 is revised slightly:

THEOREM 3′: *The change in direct and indirect tax liability that results from a marginal increase in labor income (and hence a marginal increase in expenditure) is nil for the person of lowest ability and highest ability; formally, if* $n = N_1$ *or* N_2, *then*

$$0 = \frac{\partial}{\partial z}\left[T(z) + \sum_i Q^i(x_i) + \sum_j \tau_j m_j(\cdot)\right]$$

where $\tau = q - p$. (See the Appendix for proof.)

The reader will recall from the heuristic explanation in Section I that it is inefficient to discourage these people from supplying an additional unit of labor, which is why they face zero marginal tax rates. In the multicommodity case, this means that the marginal influence of direct and indirect taxation must be nil.

An implication of Theorem 3′ is that the *direct* tax liability from a marginal increase in labor income for the person of lowest or highest ability must be negative if the *indirect* liability is positive. In other words, positive commodity taxes at the optimum imply a subsidy on marginal income from labor for the worst-off person and the best-off person. In view of this observation, I cannot expect to prove that the optimal marginal income tax is nonnegative everywhere; instead I must frame our nonnegativity proposition in terms of the *total* tax liability from marginal earnings:

THEOREM 4′: *The change in direct and indirect tax liability,* $\partial[\cdot]/\partial z$, *that results from a marginal increase in labor income is nonnegative if the following conditions are met:*

(i) *Goods subject to tax rates and leisure have nonnegative cross partials:* $\partial u_1/\partial c_j \geq 0$ *all j.*

(ii) *The marginal social value of a dollar decreases as utility increases:* $d(G'\Psi)/dv \leq 0$.

(See the Appendix for proof.) Assumption (ii) is appealing because it is fundamental to the ability-to-pay tradition, in which the tax system is used to equalize differences in utility levels.

The assumption that the utility function is additively separable between leisure and other commodities has strong implications; the following theorem is similar to a proposition by Mirrlees (1976).

THEOREM 5: *If leisure is additively separable from commodities then the change in tax liability that results from a marginal increase in use of any of these commodities is nil; formally, if* $u_{1k} = 0$ *all k, then*

$$\frac{\partial}{\partial x_k}\left(T(z) + \sum_i Q^i(x_i) + \sum_j \tau_j m_j(\cdot)\right) = 0$$

(See the Appendix for proof.)

This theorem implies, for example, that there should be no tax/subsidy on housing if there is a tax/subsidy on labor income. This result is not surprising once we recognize that the path of the state variable v_n is independent of the x's in the separable case by condition IV. Control over the path of utilities is achieved by manipulation of z; manipulation of the x's adds nothing.

III. Tax Rates for Many Commodities

I shall review Ramsey's problem and his classical conclusions before generalizing them. The Ramsey problem is to choose commodity taxes which will meet the government's revenue need at minimum loss of utility to a single, representative consumer. There are various statements of the solution; I shall offer four of them after a comment on notation.

The notation is consistent throughout the paper; Ψ is the marginal utility of expenditure and λ is the shadow price on government revenue. The ratio Ψ/λ converts private expenditure into units which are comparable to government revenue, which is taken as the unit of account. So it is useful to introduce the symbol $w = 1 - \Psi/\lambda$, which is the deviation of the social value of private expenditure from its nominal value. Thus $1 - w$ is the social value of private expenditure.

Four versions of Ramsey's rules are as follows:[11]

PROPOSITION 1: (*Cost = Benefit*) *The social cost of the reduction in private expenditure from a marginal increase in the tax*

[11]The four versions of Ramsey's rules can be derived by solving max $v(p + t) + \lambda(t'c - R)$. See Peter Diamond and Mirrlees.

rate on any commodity equals the resulting increase in government revenue; formally

$$(1 - w)m_k(\cdot) = \frac{\partial R}{\partial \tau_k} \qquad all\ k$$

PROPOSITION 2: (*Diamond and Mirrlees*) *The ratio of the tax revenue from a marginal increase in the tax on any commodity to the quantity of that commodity is constant; formally*

$$\frac{\partial R/\partial \tau_k}{c_k} = (1 - w) \qquad all\ k$$

PROPOSITION 3: (*Ramsey*) *The optimal set of commodity taxes reduces the consumption of every commodity by the same proportion; formally*

$$\sum_j \tau_j m_{jk}(\cdot)/c_k = \sum_j \tau_j \frac{\partial c_j}{\partial m} - w \qquad all\ k$$

where $m_{jk}(\cdot)$ is the Slutsky substitution term. (Notice that the right side of the equation is independent of k.)

PROPOSITION 4: (*Ramsey*) *When cross-price elasticities are nil, the optimal tax on any good is inversely proportional to its elasticity of demand; letting θ_k be the tax on value ($\theta_k p_k = \tau_k$),*

$$\theta_k = w/\eta_k \qquad all\ k$$

where η_k is the demand elasticity (defined to be positive).

There are two differences between Ramsey's rules and the propositions which I am about to derive from the general formulation. First, in Ramsey's problem the price of one commodity is varied while holding other commodity prices constant; in the general problem one control is varied while holding the others constant, which includes z and x. Some of the price effects which I shall examine do not permit a general equilibrium response by the consumer; rather he is constrained to respond by adjusting only his consumption of c_1, $c_2, \ldots$. In particular the price elasticity of government revenue in the following theorems differs from measured elasticity to the extent that consumers respond to changes in tax rates by adjusting demand for goods subject to tax schedules.

Second, in Ramsey's problem the social value of private expenditure $(1 - w)$ is the same on every good, but it is different for each good in the general problem. It is the same in Ramsey's problem because there is only one consumer and he equates the marginal utility of expenditure on different commodities in order to maximize utility. It is different in the general setting because different consumers buy the same good in different quantities and the marginal social utility of expenditure is different for different consumers. In brief the general problem must take account of the distributional consequences of commodity taxes.

Additional notation is required. Define $\overline{m}_k$ to be the nominal cost (compensating variation) of a small increase in the price of commodity c_k:

$$\overline{m}_k = \int_{N_1}^{N_2} m_k f\, dn$$

The social cost is obtained from the nominal cost by converting private expenditure into units comparable to government revenue and taking account of the distribution effect: thus $-\mu\Psi/\lambda$ is the marginal rate of substitution between m and R. It is the shadow price of an n-person's expenditure when government revenue is the numeraire. In addition m_n is the labor supply of an n-person, or, if you will, the private worth of the gross wage; the distribution effect is captured by m_{kn}, which is the effect of the price increase upon the private worth of the gross wage. Now define w_k:

$$w_k = -\int_{N_1}^{N_2} \frac{\mu\Psi}{\lambda} m_{kn}\, dn/\overline{m}_k$$

The social value of private expenditure $(1 - w)$ figured prominently in Propositions 1-4; in their generalization it is replaced by $1 - w_k$, which is the average social value of expenditure on c_k.

The generalizations of Ramsey's results are obtained from condition VIII:

$$0 = \frac{\partial J}{\partial q_k} \qquad all\ k$$

The theorems are shown in the text; the proofs are in the Appendix.

THEOREM 6a: (*Cost = Benefit*) *The social cost of the reduction in private expenditure from a marginal increase in the tax rate on any commodity equals the resulting increase in government revenue from commodities* $c_1, c_2, \ldots$; *formally*

$$(1 - w_k)\overline{m}_k = \left.\frac{\partial R}{\partial \tau_k}\right|_{xz} \qquad \text{all } k$$

THEOREM 6b: *The ratio of the tax revenue from goods* $c_1, c_2, \ldots$ *caused by a marginal increase in the tax on any commodity, to the weighted quantity of that commodity is a constant; the weight is the average social value of private expenditure on the commodity. Formally*

$$\frac{\partial R/\partial \tau_k \,|_{xz}}{(1 - w_k)\bar{c}_k} = 1 \qquad \text{all } k$$

where $\bar{c}_k$ *is total consumption of commodity k.*

THEOREM 6c: *The optimal set of commodity taxes reduces the consumption of good k by a proportion which is decreasing in the average social value of private expenditure on that good; consumption of all goods are reduced in the same proportion if the average social value of private expenditure is the same for each good. Formally,*

$$-\frac{\int_{N_1}^{N_2} \sum_j \tau_j m_{jk} f dn}{\bar{c}_k} = w_k \qquad \text{all } k$$

THEOREM 6d: *When cross-price elasticities are nil within the commodity group* c_1, $c_2, \ldots$, *the optimal commodity tax is inversely proportional to the "average" demand elasticity and decreasing in the average social value of private expenditure on the commodity; formally*

$$\theta_k = \frac{w_k}{\eta} \qquad \text{all } k$$

where $\theta_k p_k \equiv \tau_k$

$$\bar{\eta} \equiv -\int_{N_1}^{N_2} \frac{q_k}{\bar{c}_k} \left.\frac{\partial c_k}{\partial q_k}\right|_v f dn$$

We can obtain corollaries to our theorems by using the following facts, which are derived from the definition of w_k:

(i) $(m_{nk} = 0) \Rightarrow (w_k = 0)$
(ii) $(\mu < 0 \;\&\; m_{nk} > 0) \Rightarrow (w_k > 0)$
(iii) $(\mu < 0 \;\&\; m_{nk} < 0) \Rightarrow (w_k < 0)$

For example, we obtain the following corollary by plugging these facts into Theorem 6d:

COROLLARY: *Assume that cross-price elasticities are nil within the commodity group* $c_1, c_2 \ldots$ *The optimal tax rate on* c_k *is* (i) *nil if the consumption of* c_k *remains constant when earning ability n increases,* (ii) *positive when consumption of* c_k *increases with earning ability n and we have the standard case where* $\mu \leq 0$, *and* (iii) *negative when consumption of* c_k *decreases with earning ability n and we have the standard case where* $\mu \leq 0$.

(i) tells us that tax rates are inferior policy tools when their use has no distributional impact; we should use tax schedules. (ii) and (iii) tell us that commodity taxes should reduce differences in utility levels.

IV. Concluding Remarks

An objection to the optimal income tax literature is that it tells nothing about a general system of taxation. We have seen how to remedy this complaint by deriving the major theorems in a setting with many persons and many commodities. Intuitions about tax schedules obtained from the Edgeworth-Pigou tradition are misleading; that tradition leads us to expect that marginal tax liability from labor income will rise everywhere, but we find that it rises at first and later falls. The conclusion is independent of the particular value of the elasticity of labor supply, provided that its sign is not perverse or zero. The Ramsey results on commodity tax rates hold up in the general setting after adjustments are made for distributional effects of expenditure on different goods.

There remains another objection to this literature, namely that it incorporates only one kind of incentive effect from taxation; a large object is balancing precariously on a small pedestal. In particular it tells us nothing about the incentive effects from taxation

of income from capital. This criticism misses the mark insofar as income from capital is obtained in proportion to an expenditure of effort, as with human capital. It also misses the mark if capital accumulation is a consequence of saving, since the elements of x may be dated commodities. However, rapid capital accumulation by an individual is typically a consequence of superior information, either about techniques (innovation) or markets (insiders). Our model does not capture the effects of taxation upon the creation and distribution of information; this reflects the unsatisfactory state or absense of a general theory of an economy in which information is costly.

APPENDIX

Lemma used in proving Theorem 4:

If $d(G'u_1)/dn \leq 0$, as assumed in Theorem 4, then $\mu \leq 0$.

PROOF:

(i) By II we have

$$-D\mu(n) = [(G' + \lambda g_1(\cdot))f] + [\frac{\partial}{\partial v}(\frac{\partial u}{\partial n})]\mu;$$

define a and b so that

$$= [a(n)] + [b(n)]\mu$$

(ii) Solve the linear differential equation:

$$-\mu(n^*) = \int_{N_1}^{n^*} a(m)\exp\left\{\int_{n^*}^{m} b(\hat{m})d\hat{m}\right\}dm$$

The exponential function is nonnegative, so the sign at each point along the integral is given by sign of $a(m)$.

(iii) $g_1(\cdot) = -n/u_1$ by expanding the derivative

$$\therefore a(m) = \left(G'\frac{u_1}{m} - \lambda\right)\left(\frac{fm}{u_1}\right)$$

$$(a(m) \geq 0) \Leftrightarrow (G'u_1/n \geq \lambda)$$

(iv) From (iii), the assumption that $G'u_1$ is monotonic decreasing, and the fact that λ is constant, I conclude that $a(m)$ changes sign no more than once and any change is from + to −. The conclusion follows from (ii) and the transversality conditions $0 = \mu(N_1) = \mu(N_2)$.

THEOREM 4: $T' \geq 0$ *if*

(i) $u_{12} \geq 0$

(ii) $d(G'u_1)/dn \leq 0$

PROOF:

(i) From proof of Theorem 3 we have

$$0 = \frac{\partial H}{\partial x} = \mu\frac{\partial}{\partial x}[u_1 g/n^2] + \lambda(g_2 - 1)f$$

and $$g_2 = \frac{1}{1 - T'}$$

$$\therefore \frac{1}{1 - T'} - 1 = \left(\frac{-\mu}{\lambda f}\right)\frac{\partial}{\partial x}[u_1\cdot g/n^2]$$

(ii) $\mu \leq 0$ by preceding Lemma. λ is a constant by III; we can see that it is positive by complementary slackness.

(iii) We need only prove that $\partial[u_1\cdot g/n^2]/\partial x > 0$ to establish the theorem. We proceed by differentiating:

$$\frac{\partial}{\partial x}(u_1 g/n^2) = (-gg_x u_{11}/n + gu_{12} + g_x u_1)/n^2$$

We know that $u_{11} < 0$ by concavity. Also g_x is the marginal rate of substitution between z and x, which is positive. The conclusion follows from the assumption $u_{12} \geq 0$.

Comment: Obviously there are weaker conditions under which $T' \geq 0$. It is sufficient if $u_{12} \geq g_x u_{11}/n - g_x u_1/g$. I have used the condition on the cross partial in the theorem because it is intelligible, not because it is general.

Lemma used in proving Theorem 4':

If $d(G'\Psi)/dv \leq 0$ as assumed in Theorem 4', then $\mu \leq 0$.

PROOF:

(i) $-D\mu = [(G' - \lambda\sum_j p_j m_{jv})f] + [\frac{\partial}{\partial v_n}(\frac{\partial u(\cdot)}{\partial n})]\mu$ by Condition II; define a and b

$$= [a(n)] + [b(n)]\mu$$

(ii) $\sum_j p_j m_{jv}(\cdot) = \sum_j \frac{p_j}{u_{c_j}}$

$$= \sum_j \frac{p_j}{q_j \Psi}$$

by the individual maximizing conditions (11)

$$\therefore\ a(m) = \left(G'\Psi - \lambda \sum_j \frac{p_j}{q_j}\right)\frac{f}{\Psi}$$

Notice that $\sum_j \frac{p_j}{q_j}$ is constant as m varies.

(iii) The proof is completed by following steps (ii)–(iv) in the proof in the preceding Lemma (used in proving Theorem 4), recalling that v is increasing in n by Theorem 1 so that $(dG'\Psi/dv \leq 0) \Rightarrow (dG'\Psi/dn \leq 0)$.

THEOREM 3′: *If* $n = N_1$ *or* N_2, *then* $\partial/\partial z$ [*tax liability*] $= 0$.

PROOF:

(i) The individual's budget constraint may be written

$$\text{tax liability} \equiv T(z) + \sum_i Q^i(x_i) + \sum_j \tau_j m_j(\cdot)$$
$$= z - \sum_i p_i x_i - \sum_j p_j m_j(\cdot)$$

where $q = p + \tau$.

(ii) Suppose that we increase labor income z by a small amount for some n-person on the optimal path. The other controls are held constant. Part of the income will be taxed and the rest will be spent. From i we have

$$\frac{\partial}{\partial z}(\text{tax liability}) = 1 - \sum_j p_j \frac{\partial m_j(\cdot)}{\partial z}$$

(iii) From Condition I we have

$$0 = \frac{\partial H}{\partial z} = \mu \frac{\partial}{\partial z}\left(\frac{\partial u}{\partial n}\right) + \lambda\left(1 - \sum_j p_j \frac{\partial m_j}{\partial z}\right) f \quad \text{for all } i$$

(iv) Combining (ii) and (iii) gives

$$\frac{\partial(\text{tax liability})}{\partial z} = \left(\frac{-\mu}{\lambda f}\right)\left\{\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial n}\right)\right\}$$

(v) By Condition VI we have $\mu(N_1) = \mu(N_2) = 0$; this fact and (iv) establishes the theorem.

THEOREM 4′: ∂[*tax liability*]$/\partial z \geq 0$ *if*
(i) $\partial u_1/\partial c_j \geq 0$ *all* j
(ii) $d(G'\Psi)/dv \leq 0$

PROOF:

(i) From the proof of Theorem 3′ we have $\partial[\text{tax liability}]/\partial z = -\mu/\lambda f\{\partial(\partial u/\partial n)/\partial z\}$.

(ii) $\mu \leq 0$ by the preceding Lemma; $\lambda > 0$; hence we only need to prove $\{\cdot\} \geq 0$.

(iii) Repeat step (iii) of the proof to Theorem 4 to get $\{\partial[\partial u/\partial n]/\partial z\} \geq 0$ by using the assumption on the sign of the cross partials.

THEOREM 5: $(u_{1k} = 0) \Rightarrow \left(\frac{\partial}{\partial x_k}\left(T + \sum_i Q^i(x_i) + \sum_j \tau_j m_j(\cdot)\right) = 0\right)$

PROOF:

(i) Following steps (i)–(iii) of the proof to Theorem 3′, we obtain

$$-\mu \frac{\partial}{\partial x_k}\left(\frac{\partial u}{\partial n}\right) = \lambda\left(-p_k - \sum_j m_{jk}\right) f \quad \text{all } k$$

$$\frac{\partial[\text{tax liability}]}{\partial x_i} = -p_k - \sum_j m_{jk}$$

Combining the above

$$\frac{\partial[\text{tax liability}]}{\partial x_i} = \frac{-\mu}{\lambda f}\frac{\partial}{\partial x_n}\left(\frac{\partial u}{\partial n}\right)$$

(ii) $\left[\frac{\partial u}{\partial n}\right] = u_1 z/n^2$ by expanding the derivative

$$\Rightarrow \frac{\partial[\cdot]}{\partial x_k} = 0 \text{ assuming } u_{1k} = 0 \quad \text{all } k$$

(iii) Theorem 5 follows from (i) and (ii).

Lemma invoked in Theorem 6a:

$$\partial[\partial u/\partial n]\partial q = -\Psi m_{nk}$$

PROOF: (see Mirrlees, 1976)

(i) Define the indirect utility function u^*:

$$u^*(-z/n, x, q, m(q, -z/n, x, u)) \equiv \max\,[u(-z/n, x, c): \Sigma q_j c_j \le m]$$

(ii) $\frac{\partial}{\partial q_k}\left(\frac{\partial u}{\partial n}\right) = \frac{\partial}{\partial q_k}(Dv)$

by (12) and condition IV

$$= \frac{\partial}{\partial q_k}\left(\frac{\partial u^*}{\partial n}\right)$$

$$= \frac{z}{n^2}\left(\frac{\partial}{\partial q_k}\frac{\partial u^*}{\partial(-z/n)}\bigg|_m + \frac{\partial}{\partial m}\left(\frac{\partial u^*}{\partial(-z/n)}\right)\frac{\partial m}{\partial q_k}\right)$$

Rewrite the preceding expression in simpler notation:

$$= \frac{z}{n^2}\left(u^*_{q_k 1} + u^*_{m1} m_k\right)$$

(iii) Roy's Rule: $u^*_{q_k} = -\Psi m_k$

(iv) Combine (ii) and (iii).

$$\frac{\partial}{\partial q}\left[\frac{\partial u}{\partial n}\right] = \frac{z}{n^2}\left[\frac{\partial}{\partial(-z/n)}(-\Psi c_{q_k}) + u^*_{1m} m_k\right] \text{ for } v \text{ constant}$$

$$= \frac{z}{n^2}\left[-\Psi\frac{\partial(m_k(\cdot))}{\partial(-z/n)} - m_k u^*_{1m} + u^*_{1m} m_k\right]$$

$$= -\Psi m_{nk}$$

using $m_q = c_q$ and $\partial\Psi/\partial(-z/n) = u^*_{m1} = u^*_{1m}$

THEOREM 6a: $(1 - w_k)\overline{m}_k = \frac{\partial R}{\partial \tau_k}\bigg|_{xz}$

PROOF:

(i) By VIII we have

$$0 = \frac{\partial J}{\partial q_k} = \int_{N_1}^{N_2}\left\{\mu\frac{\partial}{\partial q_k}\left(\frac{\partial u}{\partial n}\right) - \lambda\sum_j p_j m_{kj} f\right\} dn \quad \text{all } k$$

(ii) By the preceding Lemma, we have:

$$\frac{\partial}{\partial q_k}\left(\frac{\partial u}{\partial n}\right) = -\Psi m_{nk}$$

(iii) $\Sigma_j p_j m_{kj} = \Sigma_j (q_j - \tau_j) m_{jk}$ since $p_j = q_j - \tau_j$ all j

but $\sum_j q_j m_{jk} = 0$ (See John Hicks, p. 311.)

$$\therefore \sum_j p_j m_{kj} = -\sum_j \tau_j m_{kj}$$

(iv) The change in tax revenue is

$$\frac{\partial R}{\partial q_k}\bigg|_{xz} = \int_{N_1}^{N_2}\frac{\partial}{\partial q_k}\left[T(z) + \sum_i Q^i(x_i) + \sum_j \tau_j m_j\right] f\,dn$$

where $[\cdot] \equiv$ tax liability by definition

$$= \int_{N_1}^{N_2}\left(m_k + \sum_j \tau_j m_{jk}\right) f\,dn$$

(v) Combine (i), (ii), (iii), and (iv) to get

$$0 = -\int_{N_1}^{N_2}\frac{\mu\Psi}{\lambda} m_{kn}\,dn - \int_{N_1}^{N_2} m_k f\,dn + \frac{\partial R}{\partial q_k}\bigg|_{xz}$$

The theorem follows from v and our definitions of $\overline{m}_k$ and w_k.

THEOREM 6b: $\frac{\partial R/\partial\tau_k|_{xz}}{(1 - w_k)\bar{c}_k} = 1$

PROOF:

Rearrange terms in Theorem 6a and use the fact that

$$\frac{\partial m(\cdot)}{\partial q_k} = c_k \Longrightarrow \int_{N_1}^{N_2} m_k(\cdot) f\,dn = \int_{N_1}^{N_2} c_k f\,dn = \bar{c}_k$$

THEOREM 6c:

$$-\frac{\int_{N_1}^{N_2}\sum_j \tau_j m_{jk} f\,dn}{\bar{c}_k} = 1 - (1 - w_k)$$

PROOF:

(i) Combine steps (i)–(iii) in the proof of Theorem 6a to obtain

$$0 = \int_{N_1}^{N_2}\left(-\mu\psi m_{nk} + \lambda\sum_j \tau_j m_{kj} f\right) dn$$

(ii) By symmetry of the Slutsky matrix we have $m_{jk} = m_{kj}$.

(iii) By definition

$$w_k = -\int_{N_1}^{N_2} \frac{\mu\Psi}{\lambda} m_{kn} dn / \overline{m}_k$$

$$\overline{m}_k = \bar{c}_k$$

$$\therefore w_k \bar{c}_k = -\int_{N_1}^{N_2} \frac{\mu\Psi}{\lambda} m_{kn} dn$$

(iv) Combine (i), (ii), and (iii) to get Theorem 6c.

THEOREM 6d: $\theta_k = \frac{w_k}{\bar{\eta}}$

PROOF:

(i) Rewrite Theorem 6c on the assumption that cross-price effects are nil:

$$-\frac{\int_{N_1}^{N_2} \tau_k m_{kk} f\, dn}{\bar{c}_k} = w_k$$

(ii) Define

$\theta_k \equiv \tau_k / p_k$, the expenditure tax on good k

$\bar{\eta} \equiv -\int_{N_1}^{N_2} \left(\frac{q_k}{\bar{c}_k} \frac{\partial c_k}{\partial q_k}\Big|_v f \right) dn$, a measure of average elasticity of compensated demand.

(iii) Assume in the pretax situation $p_k = q_k$; use this fact and (i) and (ii) to obtain Theorem 6d.

REFERENCES

Kenneth J. Arrow and Mordecai Kurz, *Public Investment, the Rate of Return and Optimal Fiscal Policy*, Washington 1972.

R. Cooter and E. Helpman, "Optimal Income Taxation for Transfer Payments," *Quart. J. Econ.*, Nov. 1974, *88*, 656–70.

R. Dorfman, "An Economic Interpretation of Optimal Control Theory," *Amer. Econ. Rev.*, Dec. 1969, *59*, 817–31.

P. Diamond and J. Mirrlees, "Optimal Taxation and Public Production," *Amer. Econ. Rev.*, Part I Mar. 1971, *61*, 8–27; Part II, June 1971, *61*, 261–78.

F. Y. Edgeworth, *Papers Relating to Political Economy*, London 1925.

R. E. Fair, "The Optimal Distribution of Income," *Quart. J. Econ*, Nov. 1971, *85*, 551–79.

M. Feldstein, "On the Optimal Progressivity of the Income Tax," *J. Publ. Econ.*, Nov. 1973, *2*, 357–76.

Harold M. Groves, *Tax Philosophers*, Madison 1974.

John R. Hicks, *Value and Capital*, Oxford 1968.

J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 175–208.

———, "Optimal Tax Theory: A Synthesis," *J. Publ. Econ.*, Nov. 1976, *6*, 327–58.

E. S. Phelps, "Taxation of Wage Income for Economic Justice," *Quart. J. Econ.*, Aug. 1973, *87*, 331–54; reprinted in his *Economic Justice*, New York 1973, 417–38.

E. Sadka, "On Income Distribution, Incentive Effects and Optimal Income Taxation," *Rev. Econ. Stud.*, forthcoming.

J. Seade, "Progressivity of Income Taxation," mimeo., Oxford 1975.

E. Sheshinski, "The Optimal Linear Income Tax," *Rev. Econ. Stud.*, July 1972, *39*, 297–302; reprinted in Edmund S. Phelps, ed., *Economic Justice*, New York 1973, 409–16.

Fixed Rules vs. Activism in the Conduct of Monetary Policy

By ROGER CRAINE, ARTHUR HAVENNER, AND JAMES BERRY*

The objective of this paper is to examine the issue of fixed rules vs. "activism" in the conduct of monetary policy. A variety of policies ranging from simple fixed growth rules to elaborate optimal control schemes have been urged on monetary policymakers for years, and there has been considerable debate by the proponents about the conditions for superiority of the proposed schemes. Arthur Okun (1972) criticizes those favoring fixed rules as depending on 1) inherently stable private demand, 2) costs to policy changes, and 3) relations between targets and controls which are known in the long run, but unknown in the short run. Milton Friedman seems to agree in principle. In the course of debate with Walter Heller, he stated, "if I thought I could predict precisely, well then,..., I would be prepared to make fine adjustments to offset other forces making for change" (pp. 49–50). He justifies fixed rules by emphasizing the lag in recognizing the need for policy change and uncertainty in predicting correctly what will result from the changes in the short run.

The view that best characterizes the world must be determined empirically and there have been numerous attempts; however, there are many pitfalls in assessing alternative policies. Evaluation of the policies necessarily requires some quantitative statement about the unknown true economic structure and some criterion to judge which policy is "best." If the approximation to the economic structure is too sensitive to changes in the policy variables, then the evaluation will not be valid. Several early attempts (see, for example, Martin Bronfenbrenner and Franco Modigliani) used the time derivative of the equation of exchange ($MV = PY$) to evaluate the policies. This structure implies that the policy impact occurs within the period and changes in velocity and/or real output are independent of changes in the policy. Later work used econometric models (see, for example, Jan Kmenta and Paul Smith; J. Phillip Cooper and Stanley Fischer; George Perry, 1975a) which are much less restrictive since all the endogenous variables may be affected by the policy changes and there may be lagged effects. Still, for tests of historical periods it must be assumed that the error realizations are independent of the policy settings. In addition to an economic structure, an information structure must be assumed. Activists' policies should be based on all the information available to the decision maker *ex ante*, such as forecasts of the exogenous variables, judgmental information on current or future error realizations, and any other currently observable data, but not on information which is available only *ex post*, such as currently unobservable variables, or actual future values of the exogenous variables or error realizations. Finally, comparing actual policy to alternative rules (as Kmenta and Smith and Perry did) uses the correct information, but there is of course no assurance that any tested policy was in fact optimal for the given model. Cooper and Fischer searched for an optimal policy, but "... because of the formidable computational difficulties of finding optimum controls in a large nonlinear dynamic model like the *FMP* [FRB-MIT-Penn] model" (p. 750),[1] the selected policies were constrained to a particular

*University of California-Berkeley, New York University, and Federal Reserve Board, respectively. The views expressed herein are solely our own and do not necessarily represent the views of the Board of Governors of the Federal Reserve. We would like to thank George Borts, an anonymous referee, and members of the Board's staff for helpful comments

[1]The *FMP* model is now called the MIT-Penn-SSRC or *MPS* model.

(suboptimal) form of reaction function, and they did not take advantage of judgmental forecast information.

This paper is an attempt to remedy some of these deficiencies by examining the performance of a variety of policies (including a formally optimal policy based on only current information) in as realistic a setting as possible.[2] The policies are evaluated over the extremely volatile two-year period from 1973-III–1975-II. The period includes the highest unemployment and inflation rates experienced since World War II and provides a stringent test for any policy. There were large unpredictable exogenous shocks—which should favor active policy—coupled with increasing uncertainty about the structure of the economy, which should favor *ex ante* rules. We chose the large (180 equation) non-linear structural *MPS* econometric model[3] as a description of the economy, because it has been used in numerous policy and forecasting experiments in the past and has a well-developed monetary sector.[4]

The problem is to select the money stock sequence, given only the information available in the quarter the decision is made, that minimizes *ex post* a loss function that penalizes unemployment[5] rates (u) in excess of 4.8 percent, inflation[6] rates ($\dot{p}$) greater than 2.5 percent, changes in the Treasury Bill rate (Δr_t^{TB}) of more than 150 basis points, and deviations in M_1 from a 5.1 percent growth path subject to being consistent with the *MPS* quarterly economic model. With h the horizon (in quarters), the loss function is:

$$(1)\quad L = \sum_{t=1}^{h} 2(u_t > 4.8)^2 + 1(\dot{p}_t > 2.5)^2 + 5(|\Delta r_t^{TB}| > 1.5)^2 + .0001(M_{1t} - 1.051^{t/4} M_{1o})^2$$

where u = unemployment
$\dot{p}$ = inflation rate
$|\Delta r^{TB}|$ = absolute value of the change in the Treasury Bill rate
M_1 = demand deposits and currency, where M_{1o} refers to 1973-II.

The target paths in the loss function for inflation (2.5 percent) and unemployment (4.8 percent) were chosen based on the Nixon Administration's announced objectives in 1973. Their goal was to reduce the inflation rate to the 2–3 percent range while maintaining full employment.[7] In the *MPS* model, the long-run "natural" rate of unemployment is 4.8 percent with prices increasing at the rate of increase of the money stock less 2.6 percent. Two monetary targets are also included. Although the Federal Reserve did not publicly announce its desired money growth path until March 1975, they were following an aggregates policy subject to smoothing short-term interest rate fluctuations in 1973 (see the explanation of actual policy below).

The targeted 5.1 percent growth rate for the money stock is within the range made public in 1975, and is consistent with the inflation and unemployment target paths, that is, the loss function is zero in nonstochastic steady-state equilibrium. The discontinuous penalty on Δr^{TB} penalizes large quarterly changes (>1.5 percent) in the Treasury Bill rate since it is felt that large fluctuations in short-term rates increase uncertainty in financial markets.

The loss function is evaluated by dynamically simulating the various policies on the *MPS* model with the exogenous variables[8] and error realizations set at their

[2] Current computational restrictions still limit us to computing first-order certainty equivalence control policies.

[3] See the *Quarterly Econometric Model Data Directory* and the *Quarterly Econometric Model Equations* for details.

[4] See Carl Christ; Cooper; Cooper and Charles Nelson; Edward Gramlich; Thomas Muench et al.; Perry (1975a); James Pierce and Jared Enzler.

[5] Unemployment rate of total labor force (includes armed forces).

[6] Price deflator for nonfarm business product, excluding household, net of federal indirect business taxes.

[7] For example, see the *Economic Report of the President*, p. 82.

[8] Discretionary fiscal policy may be a function of monetary policy (as the referee pointed out) and therefore not truly exogenous (the automatic stabilizers are

historical values. The simulated values of the endogenous variables are used to evaluate the loss.

The remainder of the paper is organized as follows. Section I presents the policies to be tested, Section II gives the results and an analysis of the policy breakdowns, Section III tests the sensitivity of the results to the loss function weights and the desired paths, and Section IV is the conclusion.

I. Policies

Six policies are constrasted over the period. Two of the policies, labled Friedman and Poole, *do not* make use of information as it becomes available through the control horizon and are called fixed rules, being given *ex ante*. Four other policies, labeled Actual, Bronfenbrenner, Cooper-Fischer, and Feedback Optimal, make use of selective information (see the policy explanations below) as it becomes available. These policies are denoted "active" since they are modified during the control period (*ex durante*), although the kind and quantity of economic information used by the policies is notably different.

A. *Friedman-Type Fixed Rule (ex ante)*

Friedman's constant money growth arguments are by now well known to economists. A fixed growth rate policy is independent of initial conditions, specific knowledge of the structure of the economy, and anticipated exogenous events. We chose a fixed money growth rate of 5.1 percent which is close to Friedman's own recommendation for the period (see Friedman, p. 3), and minimizes the loss function (equation (1)) in a deterministic steady-state equilibrium.

B. *Poole's Cautious Reentry (ex ante)*

William Poole argued that "what is needed is a measure of the level of the money stock relative to the established trend" since

> Once the rate of inflation has adjusted to a given rate of money growth, a change in the rate of money growth can be expected to produce an effect on aggregate activity. However, it surely must be the case that a five percentage point drop in the rate of money growth will produce a cycle peak sooner than a one percentage point drop in the rate of money growth, other things equal. [1973, p. 6]

He proposed a two-stage rule in which money growth proceeds at 4 percent until the money stock reaches 98 percent of an established trend, at which time money is increased at 6 percent until "the pause in economic activity is clearly over" (1973, p. 26).[9] The trend is defined by the highest (not average) monthly money growth rate over a twenty-four month period in the neighborhood of the cycle peak, running the trend through the average level of the money stock over the particular period that had the highest rate (Poole, 1973, p. 8). (The somewhat exaggerated hypothetical example of Figure 1 may aid in understanding Poole's rule.) Over our period of 1973-III-1975-II, this implies 4 percent money growth until 1974-II when 6 percent becomes the recommended policy.[10]

C. *Actual Monetary Policy (ex durante)*

According to the Federal Reserve *Bulletins* of the period,[11] open market operations were determined by setting fixed growth rules (from last quarter's base) for

already endogenous). For this time period, however, government spending—the major countercyclical fiscal variable in the *MPS* model—does not appear to have been used as a countercyclical tool nor to have responded to monetary policy. Most forecasters overestimated government spending (see Okun, 1975) and indeed actual government spending in 1973 and 1974 fell short of targeted government expenditures.

[9] As Poole has observed in private correspondence, his proposal is actually a hybrid rule/discretion recommendation. The words "clearly over" are obviously ambiguous (requiring discretion). This is not a problem over our control horizon, but might be in other periods.

[10] Based on seasonally adjusted monthly data.

[11] See the "Record of Policy Actions..." sections of the *Bulletins*

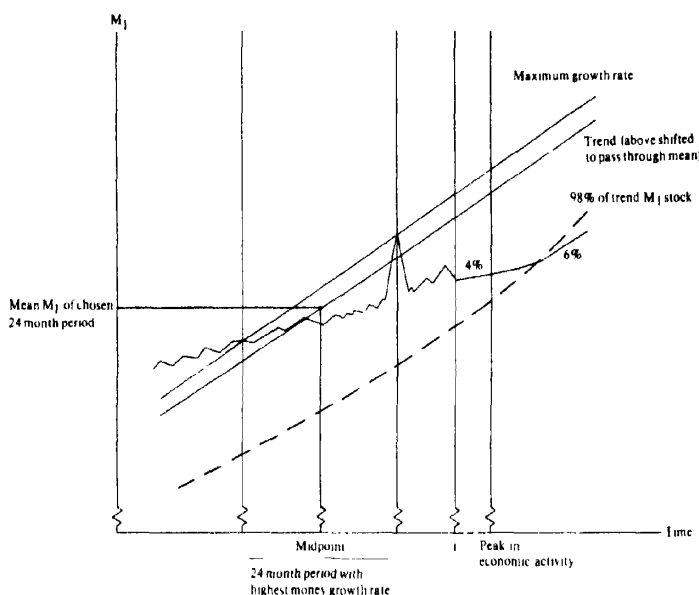



FIGURE 1. POOLE'S MONEY GROWTH RULE

At time t, M_1 is to grow at 4 percent until it reaches (dashed line) 98 percent of a trend line defined as the maximum twenty-four month growth rate in the neighborhood (labeled Maximum growth rate) shifted to pass through the average of the twenty-four months (labeled Trend), when M_1 is to revert to 6 percent until "the pause in economic activity is clearly over" (Poole, 1973, p. 26)

monetary aggregates subject to a Federal funds rate constraint. If point estimates of the aggregates and funds rate were specified, this policy scheme would be consistent with[12] the combination policy of Poole (1970). In practice, since policy is set daily and even hourly, the Federal Reserve uses tolerance ranges to filter out minor short-run variations. For example, the Fed has argued

> There is little reason to permit sharp short-run swings in interest rates (for example, 4 or 5 percentage points over a month or so) in an effort to smooth out temporary variations in money and credit. [1974, pp. 336 37]...
> The inherent short-run volatility of the monetary aggregates is one reason why the Committee expresses its short-run guides in terms of ranges of tolerance.... This may reflect transitory factors that are influencing money but are expected to be self-correcting, as for instance, when a sharp drop in U.S. Government deposits results in temporary bulges in private demand balances before the funds are invested in other assets by holders. [1974, p. 334]

Thus, while the combination policy optimizes over truly random realizations of the structural equation errors, inconsistencies in the prescribed ranges of the aggregates or the federal funds rate are assumed to be indicative of nontransitory external shocks, requiring policy action. Once the federal funds rate or an aggregate falls outside the tolerance range, the probability that it represents a structural shift rather than a simple outlier increases. Any information the staff can supply on where the structural shift occurred is then critically important, since the appropriate policy response is reversed depending on whether the *IS* or

[12]Although not necessarily the same as the combination policy The relevant question is whether the experience of the policymaker results in the optimal combination.

LM curve moved.[13] Thus, actual monetary policy can be viewed as an optimal combination policy subject to modification dependent on judgmental information about possible structural changes.

D. *Bronfenbrenner's Lag Rule (ex durante)*

In 1961, Bronfenbrenner (1961a,b) tested a proposition he called the lag rule, in which the money supply each quarter is increased by last quarter's percentage increase in productivity per man-hour, percentage increase in the labor force, and percentage decrease in velocity. Bronfenbrenner compared various rules to an "...'ideal' [policy]... estimated to hold the price level constant" (1961b, p. 622).[14] Rather than specifying a model and welfare function, he tested for similarities of each of the policies to the ideal and found that, "As in the longer-term study also [on annual data], the lag rule does best by a wide margin as regards algebraic fluctuations from the ideal pattern (avoidance of inflationary bias)" (1961b, p. 624). Since our loss function targets the inflation rate at 2.5 percent rather than zero, we have adjusted the lag rule by adding 2.5 percent annual money growth.

E. *Ad Hoc Feedback (ex durante)*

Cooper and Fischer used stochastic simulations of the *MPS* model to derive and evaluate four alternative feedback rules determining the growth of M_1 from rates of unemployment and inflation and changes in these rates. The rule they found to be "clearly the best"[15] was

$$\begin{aligned}\dot{M}_{1t} = .01 - .5(\dot{p}_{t-1} - .0062) \\ + .5(u_{t-1} - .0466) \\ - 2(\dot{p}_{t-1} - \dot{p}_{t-2}) + 2(u_{t-1} - u_{t-2})\end{aligned}$$

where $\dot{M}_1$ is the quarterly rate of growth of the money stock, $\dot{p}$ is the quarterly inflation rate, and u is the unemployment rate. A search procedure was used by Cooper and Fischer to determine the feedback coefficients:

> ... in particular, our "target" rates of inflation and unemployment, [.0062] and [.0466] respectively, were set at the mean rates of inflation and unemployment obtained in a simulation of a constant growth rate rule (at the rate [.01]) to insure that our choice of targets was consistent with the "Phillips curve" embodied in the model and the search was conducted over different values of the [remaining parameters, excluding the intercept of .01]. [p. 751]
>
> One rule was preferred to another if it produced a lower standard deviation in either of the target variables without adversely affecting the other target. [p. 750]

F. *Feedback Optimal (ex durante)*

Since the publication of Cooper and Fischer's work, improved computing facilities have made it possible to find the sequence of money stock changes that minimizes the loss function, a policy derived by optimization that is necessarily the best policy in the absence of uncertainty.[16] For a realistic test, however, the policy must be based on

[13]This argument follows directly from Poole (1970). Assuming for expositional simplicity that all targets are adequately approximated by the level of income Y, then Poole's model is appropriate.

$$Y = a_0 - a_1 r + u$$
$$M = b_0 + b_1 Y - b_2 r + v$$

where r is an interest rate and M a monetary aggregate (all coefficients > 0) If a structural shift sets $E(u) \neq 0$ (moves the *IS* curve), then the appropriate policy response is an offsetting movement in r. On the other hand, if $E(v) \neq 0$, the preferred response is to allow M to change (since Y is unaffected).

[14]Differentiating the equation of exchange ($MV = PY$, where V is velocity and Y is real output) with respect to time gives $\dot{M} + \dot{V} = \dot{P} + \dot{Y}$. If velocity and real output are invariant with respect to policy changes, then the "ideal" policy is $\dot{M} = -\dot{P}$.

[15]See equation (1.1), p. 757. (There are two typographical errors in this equation as printed.)

[16]How much better, how sensitive to timing nuances, and the possibilities of reaching optimal aggressive policy by successive approximations from more timid policy have been examined in Craine, Havenner, and Peter Tinsley.

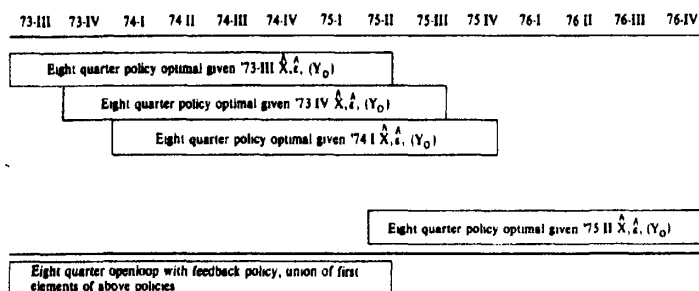



FIGURE 2. OPENLOOP WITH FEEDBACK

$\hat{X}_t$ = matrix of forecase (at time t) exogenous variables for the next eight quarters
$\hat{\epsilon}_t$ = matrix of forecast (at time t) residuals for the next eight quarters
$(Y_0)_t$ = matrix of initial conditions at time t

only the information available in the quarter in which the policy is being determined.

In each quarter, an optimal monetary policy can be calculated for the succeeding eight quarters based on the eight-quarter forecast of the exogenous variables and equation residuals made at that time. The *MPS* model has 136 exogenous variables, the majority of which are easily and accurately forecast two years into the future, for example, time trends, age groups of the population, tax rates, strike dummies, etc. A small subset of variables are much more difficult to forecast. These include exports, import prices, and discretionary federal government spending.[17] Given an eight-quarter optimal control solution based on the current forecasts, the model can be simulated for one quarter, using the first quarter of the optimal policy sequence and the historical exogenous variables and errors, to evaluate the endogenous variables. At the beginning of the next quarter, new information (error and exogenous variable realizations for the preceeding quarter and updated forecasts of the future) becomes available and a new eight-quarter solution can be calculated based on the new information. Simulating this policy for a single quarter and then repeating the entire process for each of the remaining quarters in the horizon results in a procedure we call openloop (calculation of the multiperiod solution rather than a linearized feedback rule) with feedback (modification of the solution as new information becomes available). Figure 2 shows the process schematically.

II. Results

The results of stochastic[18] simulations based on each of the six policies are given in Table 1, as well as a perfect foresight solution[19] to use as a benchmark. The second line of the table ranks the policies by minimum loss; somewhat to our surprise, it is clear that policies that used a minimal amount of information gave the best performance.

Both Friedman's fixed money growth rule and Poole's cautious reentry incurred less loss than the Cooper-Fischer *ad hoc* feedback rule and the computationally intensive feedback optimal policy (see line 1 of Table 1). Bronfenbrenner's active policy was the worst, and actual policy was slightly better than the fixed rules.

Normally feedback policies can be ex-

[17]The automatic stabilizer portions of the government budget are endogenous—tax receipts and unemployment transfers depend on endogenously generated income and unemployment.

[18]The historical residuals were added in as the structural equation errors in each period as each policy was simulated.

[19]The perfect foresight solution is based on the actual exogenous variables, using historical errors as estimates of the equation residuals; obviously, it is only possible *ex post*.

TABLE 1—TARGETS, INSTRUMENTS, AND LOSS FOR ALTERNATIVE POLICIES[a]

	Ex Ante		*Ex Durante*				*Ex Post*
	Fixed 5.1% Money Growth (Friedman)	4-6% Reentry (Poole)	Actual Monetary Policy	Lag Rule (Bronfenbrenner)	*Ad Hoc* Feedback (Cooper-Fisher)	Feedback Optimal	Perfect Foresight[d]
Loss Function Values	619.9	617.0	613.1	2336.5	803.2	677.8	600.8
Rank	3	2	1	6	5	4	
Instrument, M_1[b]							
1973-III	5.1	4	0.72	6.87	−0.67	3.49	0.47
IV	5.1	4	7.88	0.70	0.25	2.72	1.60
1974-I	5.1	4	7.10	−4.98	−2.67	5.87	8.19
II	5.1	6	6.47	−10.54	−0.89	1.72	11.74
III	5.1	6	1.36	−12.61	4.59	8.33	6.01
IV	5.1	6	3.04	−14.42	14.53	15.89	8.28
1975-I	5.1	6	4.77	−14.87	15.42	19.04	25.59
II	5.1	6	11.11	−8.35	28.43	4.49	31.82
Targets							
Inflation Rate[c]							
1973-III	5.77	5.75	5.72	5.81	5.70	5.75	5.71
IV	8.62	8.56	8.48	8.79	8.33	8.53	8.40
1974-I	12.81	12.69	12.63	13.07	12.24	12.64	12.45
II	13.89	13.72	13.74	14.07	13.00	13.62	13.47
III	12.31	12.13	12.21	12.15	11.33	12.01	12.00
IV	12.71	12.61	12.66	12.26	12.02	12.54	12.67
1975-I	11.52	11.50	11.46	10.91	11.11	11.60	11.79
II	3.75	3.80	3.65	2.66	3.50	4.29	4.74
Treasury Bill Rate Change							
1973-III	0.296	0.623	1.72	−0.198	2.261	0.779	1.811
IV	0.790	1.031	−0.82	2.584	2.025	1.452	1.498
1974-I	0.678	0.793	0.12	4.880	2.923	−0.174	−1.525
II	0.875	0.162	0.53	10.542	1.499	2.253	−1.577
III	−1.876	−2.364	0.04	10.924	−4.631	−4.515	−1.649
IV	−1.162	−1.386	−0.83	11.113	−5.894	−3.572	−1.480
1975-I	−1.123	−0.966	−1.61	−6 305	−1.856	−0.815	−1.655
II	1.156	1.083	−0.36	−4.077	−1.150	1.669	−0.309
Unemployment Rate							
1973-III	4.57	4.58	4.64	4.51	4.66	4.59	4.64
IV	4.46	4.52	4.63	4.25	4.81	4.56	4.72
1974-I	4.79	4.93	5.02	4.47	5.56	5.00	5.27
II	4.80	5.02	5.02	4.57	6.11	5.15	5.38
III	5.20	5.44	5.38	5.37	6.91	5.63	5.72
IV	6.27	6.50	6.44	7.04	8.13	6.66	6.62
1975-I	7.98	8.17	8.15	9.32	9.87	8.20	8.05
II	8.47	8.62	8.66	10.36	10.39	8.43	7.99

[a]Evaluated given actual values of the exogenous variables, with historical residuals added.

[b]M_1 is the money stock growth rate at annual rates. The base, M_1(1973-II), was $265.70 billion.

[c]Quarterly values at annual rates, nonfarm business product excluding households, net of federal indirect business taxes.

[d]Based on the actual exogenous variables, using the historical errors as *ex ante* estimates of the equation residuals.

pected to give better results than fixed rules since observed errors can be offset rather than having a lasting influence as they feed through the system dynamics. However, given uncertainty about the structure of the economy and future events there is no guarantee that performance will be improved, as the two feedback policies dem-

TABLE 2—COOPER-FISCHER *Ad Hoc* FEEDBACK MONEY GROWTH DECOMPOSITION

Quarter	M_1 Growth[a]	Fixed Term	Price Terms	Unemployment Terms
1973-III	-.0016	.01	-.0107	-.0009
IV	.0006	.01	-.0068	-.0026
1974-I	-.0067	.01	-.0205	.0038
II	-.0022	.01	-.0317	.0195
III	.0113	.01	-.0169	.0182
IV	.0345	.01	-.0027	.0272
1975-I	.0365	.01	-.0154	.0417
II	.0646	.01	-.0063	.0608

[a]Quarterly rates; Table 1 gives these same figures as quarterly values at annual rates.

onstrate. The *ad hoc* feedback rule reacts as static analysis would indicate: initially moving to offset what later turned out to be relatively low inflation rates (5.81 and 8.79 percent in the first two quarters), and reacting very strongly to the 1974-I inflation rate of 12.24 percent (simulated, given the Cooper-Fischer policy; actual was 12.63 percent), the rule severely curtailed money growth until far too late. The price terms in the rule dominate through the first three quarters, and the unemployment terms do not gain control until the last three quarters (beginning 1974-IV), (see Table 2). By this time the high unemployment rates of 8.13, 9.87, and 10.39 percent are inevitable.

The formally optimal policy (feedback optimal) did markedly better since it considered the future impact of current policy and anticipated future exogenous events. The major failure of the feedback policy occurred because the policy was too restrictive in the first year, due to much higher anticipated real sector growth than actually occurred when government spending and exports fell short of their forecasts[20] and imports far exceeded the forecast value. The most damaging error, however—although it seems counterintuitive at first—was excessively restrictive policy because of serious *under*estimates of future inflation.[21] It is now widely claimed (see Robert J. Gordon or Okun 1975), that a major portion of the inflation was due to supply-side shocks and was largely exogenous. In the *MPS* model, these shocks entered as a once-and-for-all change in the price level, but had little anticipated future inflationary effect. The optimal policy, given an exogenous sectoral price shock, depends on whether the system has flexible or inflexible prices. If the policy is restrictive but prices flexible, then the system will adjust (although aggregate demand may fall) and full employment is assured. On the other hand, if prices are sticky then the price shock has a multiplier effect which reduces aggregate demand and employment if not offset by expansive policy (see Gordon for a detailed analysis). Prices are inflexible in the short run in the *MPS* model; if the exogenous shock that increased the price level had been properly forecast, the first-year policy would have been more expansive in order to maintain the level of real balances and aggregate demand. This is apparent from the perfect foresight solution where money growth averaged 5.5 percent for the first year and accelerated rapidly in the beginning of 1974 to accommodate the properly anticipated higher price level.[22] In contrast.

[20]See Okun (1975).

[21]The average one-period ahead inflation forecast error was 5 percent compared to the analogous unemployment rate error of 0.25 percent.

[22]The final quarters in the perfect foresight solution (only) are suspect since the policy was calculated for a fixed two-year horizon. As a result, there is no consideration of the 1975 policy on the 1976 outcomes. The feedback optimal policy was calculated with a floating two-year horizon to minimize any endpoint effects.

the feedback optimal policy was not expansive until the second half of 1974.

The erroneous forecasts that sabotaged active control were not unique to the *MPS* model, which in fact fared as well or better than most others. *Business Week* magazine polled twenty-five business economists, and in addition reported seven model forecasts in its December 22, 1973 issue, p. 49. The mean inflation forecast was 6.1 percent and the highest forecast was 7.5 percent—in a year in which the implicit *GNP* deflator averaged more than 12 percent. In retrospect, it appears that four major[23] factors interacted to produce the forecast errors.

A. *Exogenous Shocks*

Over the control horizon two unanticipated large and interrelated shocks hit the system: the value of imports increased dramatically due to actions of the oil producers' cartel, while Soviet crop failures, midwestern *U.S.* floods, and a decrease in the Peruvian anchovy catch caused a worldwide protein shortage.[24] Hathaway concludes that the anticipated spread of the Green Revolution (leading to withdrawal of land from production) combined with an obsession with carrying costs and market depressing effects of large grain stocks, led exporting countries to deplete their stocks. They could no longer buffer shocks like those caused by weather, especially in the face of rising demand, as world population increased and spreading affluence led to substitution of animal proteins requiring several times the caloric input as the direct consumption of grain. The resulting inflationary pressure was further aggravated by the effect of the oil price increases on agricultural commodities, as well as the much discussed direct and indirect effects of the oil increases on the rest of the economy.

B. *Price Expectations*

Another source of errors in inflation forecasts lay in the formation of price expectations. If other factors cause price increases to exceed some threshold that triggers an unusual awareness of inflation,[25] new models of expectation formation will be required; price equations based on long distributed lags into the past that are adequate for moderate secular price changes cannot be expected to model expectation formation at inflation rates four times the sample period rates. Judgmental forecasts of the residuals would normally incorporate part of this information, but at least one key indicator of inflationary expectations was misleading: short-term interest rates, usually the sum of a pure time preference component and an expectations component (since capital loss is not an important factor), remained below 9 percent in 1974, while the inflation rate hovered around 12 percent. In retrospect, we speculate that these rates were probably depressed by the flight of stock market money to the bill market to avoid the tremendous uncertainty introduced by oil price increases, and the judgmentalists were left with an uninformative indicator of expectations.

C. *Speculation*

Richard Cooper and Robert Lawrence argue that increased commodity demand, due to coincident economic expansions in Europe, Japan, and the United States, combined with producers' reluctance to increase supply because of uncertainty about both environmental requirements and the permanence of the demand shift created vast raw material shortages. Since input shortages cause expensive disruption of production processes while simultaneously the existing inflation produced sig-

[23]The 1973 devaluation of the dollar might also be expected to stimulate exports and cause inflation in a period of supply constraints, but see Richard Berner et al. where it is argued that "... it is unlikely that recent (including the 1973 devaluation) changes in the effective exchange rate of the dollar have been a major cause of domestic inflation" (p 3).

[24]See especially Dale Hathaway, and Pierce and Enzler.

[25]For example, Otto Eckstein and Roger Brinner utilize an "inflationary severity factor" that enters when the average of the past two years' annual inflation rate exceeds 2.5 percent!

nificant inventory profits, it became economic to increase the size of raw material inventories, especially since any increased costs could be reclaimed in higher finished goods prices during the expansive period. As producers all tried to increase their inventories, they drove commodity prices still higher which ultimately raised both finished goods prices and inflationary expectations.

D. *Rebound from Price Controls*

Although the most rigid periods of price control preceded the period we examine, the effects of Phase III remain very much in evidence. The gap of weak control between January and April 1973 left the economy with a wholesale price increase of 21 percent in the second quarter of 1973 and precipitated Phase III in May, which pushed wholesale price increases down to 8 percent in the last quarter of 1973 before the controls were gradually lifted, industry by industry. In the third quarter of 1974 after controls were terminated on April 30, 1974, the wholesale price index climbed by a peak rate of 31 percent as businessmen exercised their new freedom. No price equation can be expected to model the return from a unique period and none did.

These factors combined to make economic forecasting over the period 1973-III–1975-II a hazardous undertaking, and none of the major forecasters did well.[26] The remaining two "active" policies—actual policy and Bronfenbrenner's lag rule, representing opposite extremes of performance —did not directly use the structural information in the *MPS* model, but did use some of the information which misled the forecasters.

Bronfenbrenner's lag rule is designed for secular rather than cyclical price variations and it did poorly when evaluated with our loss function over the most violent cycle since the Great Depression. Both labor productivity and the labor force participation rate vary procyclically, inducing Bronfenbrenner's rule to vary the money stock with the cycle. Okun has argued that the combined effect of this procyclical variation leads to approximately a 3 percent change in output for a 1 percent change in the unemployment rate. The only countercyclic component of Bronfenbrenner's rule—velocity—atypically moved procyclically due to the rapid inflation. Thus Bronfenbrenner's policy added to the real sector cycle.

In contrast to the long-run nature of the lag rule, actual policymakers, faced with tremendous uncertainty, seemed to have abandoned secular goals in favor of very short-run policy. After careful reading of the "Record of Policy Actions of the Federal Open Market Committee" sections of the Federal Reserve *Bulletins*, we characterize the intended policy of the period as a very slowly changing M_1 growth rate (usually a short-term rate of about 6 percent)[27] from whatever level of M_1 resulted from prior actions, subject to modification based on inconsistencies in the multiple targets.[28] The critical quarters contributing to the low loss (613) are the first four (1973-III–1974-II), when the actual money stock exceeded that of other strategies by a significant amount. This (we conjecture) occurred for a number of reasons. First, inconsistencies between the *FOMC*'s objectives and constraints[29] led to a 1973-

[26]See especially Table 2 in J. H. Kalchbrenner and P. A. Tinsley, where forecasts by Fair, the Wharton group, and a consensus of forecasts surveyed in *Business Forecasts* (Federal Reserve Bank of Richmond) are compared.

[27]Examining the April 1974–July 1975 *Bulletins* which cover January 1974–May 1975 open market actions we find that a 6 percent rate is included in all of the published ranges except the February 20, 1974 Federal Open Market Committee (*FOMC*) range, which had a lower bound of 6.5 percent. Purely subjective interpretation of the surrounding words reinforced this characterization.

[28]Although cynics often argue that federal funds rate bands dominate even long-term (quarterly) policy, there are several instances of these bands being shifted due to the behavior of the aggregates. See, for example, the May 17, 1974 revision to the April 15–16, 1974 *FOMC* directive (*Federal Reserve Bulletin*, July 1974, p. 500) or the January 9, 1975 revision to the December 16–17, 1974 directive (*Federal Reserve Bulletin*, January 1975, p. 88).

[29]See *Federal Reserve Bulletin*, January 1974, "Record of Policy Actions. . ." meeting of October 16, 1974, p. 29.

III M_1 growth far below the targeted rate. Second, a staff analysis suggested that the resulting high rates would dampen the demand for money in the quarter ahead, and this led to a 1973-IV policy that was more expansive (7.88 percent M_1 growth) *ex post* than was perhaps intended. Then, in the first quarter of 1974, in spite of M_1 and M_2 exceeding their prescribed ranges, "... in view of the sensitive state of financial markets and the general economic situation, the System [aimed] to maintain prevailing money market conditions for the time being."[30] Subjectively, it appears that long-term policy—necessarily based on forecasts which turned out to be unreliable and which led the other *ex durante* policies astray—was scrapped in favor of damped (to minimize financial market trauma) short-run reactions to the new initial conditions being faced.

In truth, actual monetary policy had a major advantage and a major disadvantage relative to the other policies considered here. The advantage was that it was not fixed for a quarter, but instead could be modified as new information became available. (The number of telephone meetings logged between the approximately monthly *FOMC* meetings attests to the perceived importance of this information.) The disadvantage was that M_1 was not directly under the *FOMC*'s control (so that the realized M_1 path was not the targeted path), unlike the assumption underlying the simulation of the other policies.[31] Compared to other policies using interim information, however, the principal difference was that feedback reactions appeared to be designed to ease the position policymakers found themselves in rather than to initiate major future improvements—the often maligned short horizon was appropriate in a period of gross uncertainty.

III. Sensitivity Analysis

Table 1 presents a clear picture of cautious policies dominating aggressive control; what conditions the results?

A. *The Loss Function*

The parameters of the criterion being minimized are necessarily of fundamental importance. In order to examine the sensitivity of the results to parameter changes, we calculated a number of differentials to determine the relative importance of variations in parameter values.[32] The differentials were calculated by 1) changing the target inflation rate from 2.5 to 5 percent per year; 2) changing the weight on deviations of unemployment in excess of the natural rate from 2 to 4; and 3) dropping the penalty on changes in the Treasury Bill rate greater than 150 basis points while raising the penalty on deviations of the money stock from its desired 5.1 percent path from .0001 to .005. Solutions were calculated based on forecast as well as actual values of exogenous variables and residuals to determine the differential with resoect to uncertainty. Large differences between elements imply that the optimal policy is sensitive to these specification changes.

As Table 3 shows, the effect of even very large changes in the parameters of the loss function is minor compared to the effects of uncertainty in the forecasts. For ex-

[30]See *Federal Reserve Bulletin*, April 1974, p. 278.

[31]The importance of this disadvantage is clear from the events of June–July, 1974 when early reports indicated that growth in M_1 would be within its tolerance range and growth in M_2 would be near its upper limit (see *Federal Reserve Bulletin*, September 1974, p. 662–63), but the federal funds rate was well above its upper limit and would not respond as usual to unborrowed reserves. After waiting a week in hope that the aberration was transitory, the open market manager was instructed to add whatever reserves necessary to bring the rate down to the targeted level (see Alan R. Homes). In spite of the high money growth in the initial weeks, M_1 finally showed a rate of increase of only 1.4 percent for that quarter.

[32]A typical differential would be

$$\delta L = \frac{\partial L}{\partial \tilde{M}_1} \frac{\partial \tilde{M}_1}{\partial p^*} \delta p^*$$

where δL is the loss differential resulting from the change in the optimal policy sequence $\tilde{M}_1$ that results from changing the (for example) targeted inflation rate p^* by some amount δp^*, here 2.5 percent. Similar differentials are calculated with respect to the other criterion parameters.

TABLE 3—SENSITIVITY OF THE FEEDBACK OPTIMAL POLICY TO LOSS FUNCTION PARAMETER CHANGES[a]

	Resulting Change in Original Loss Function Evaluation	
Policy Based On	Forecast Exogenous Variables and Residuals[b]	Actual Exogenous Variables and Residuals
Target Inflation Rate of 5 Percent Rather than 2.5 Percent	−9.3	−79.6
Unemployment Weight of 4 Rather than 2	−2.7	−64.7
Treasury Bill Rate Changes Free Rather than Penalized Beyond 1.5 Percent Movement Per Quarter	2337.1	−31.1
Original Loss Function Values	684.2	600.8

[a]Evaluation (with actual exogenous variables and residuals) using the original loss function of strategies based on alternative loss function parameters.

[b]Complete recalculation of the feedback optimal policy would involve eight multiperiod control solutions for each parameter change; these loss differentials are approximations based on one multiperiod solution for each parameter change, using as conditioning variables the exogenous variable and residual forecasts made in the immediately prior quarter

ample, basing the policy on a desired inflation rate of 5 percent rather than 2 percent results in a decrease in loss of 9.3,[33] compared to a decrease in loss of 79.6 when the target inflation rate is changed but uncertainty is eliminated by basing the policy on the actual (rather than forecast) exogenous variables and residuals. Doubling the weight on unemployment deviations has a similar but less pronounced effect. The surprising result is that even changes of 100 percent in target paths and weights do not alter the conclusion that fixed rules dominate all active control except actual policy in this period. Further, the constraint on changes in the Treasury Bill rate, which damped active policy, served to prevent even more serious blunders; when this constraint is removed the loss increases dramatically, even after the loss associated with the now active Treasury Bill rate is subtracted, as the policy based on incorrect forecasts becomes even more aggressively restrictive. The gradualism imposed by the constraint of financial market orderliness prevented an even more decisive (and disastrous) policy.

[33]Change in the original loss function resulting from the modified policy; the loss is reduced because the higher target rate results in a marginally less restrictive policy, desirable in light of inevitably higher price levels resulting from exogenous factors (see above).

B. *The Period*

The results are also conditioned by the period of the experiment: a period of substantial exogenous shocks not accompanied by structural change would favor activism. Gregory Chow has argued that active policies can be usefully employed in the presence of uncertain structure by examining a risk matrix calculated by evaluating the policy that is optimal in each model in other models. Given the matrix, a policy can be chosen by assigning probabilities that each model represents the world, or by a minimax criterion, or by any

of a variety of schemes. One difficulty is that the policymaker must be certain that the models examined bracket reality—a condition that certainly did not occur in the period 1973-II–1975-II.[34]

C. *The Model*

The *MPS* model is but one of many representations of the economic world, and specific conclusions are conditional on its validity. Just as the particular choice of loss function parameters has only minor effects on the results, however, it may reasonably be argued that use of alternative models will not importantly alter the conclusions—especially given the homogeneity of the forecasts of the period.

IV. Conclusion

This paper has examined the performance of six money stock control strategies in the context of the *MPS* model. Over the period 1973-III–1975-II, rules that were fixed *ex ante* (using no misinformation) performed well relative to more active policies that based discretionary action on an economic structure specified and estimated during more normal times, while actual monetary policy, characterized as cautiously using minimal feedback to improve conditions in the very short run, fared best of all. Given the money stock as the control variable, the results were found to be surprisingly insensitive to major changes in specification of the loss function, instead depending primarily on the highly inaccurate forecasts of the period. Large error realizations and extreme uncertainty make the choice of monetary instrument more critical; whether instrument choice is crucial in this case remains to be determined, but it is possible that an alternative instrument could have offset more of the exogenous price shocks while maintaining a favorable inflation-unemployment tradeoff.[35]

The importance of feedback has long been established in engineering applications ranging from controlling an assembly line to automatically piloting aircraft, but in this case the economic trauma inherent in the period examined caused major structural shifts roughly equivalent to a wing falling off. The principal failure was a persistent downward bias in forecast inflation, an error in which (as was pointed out) the Federal Reserve model group was not alone. Further, if the problem was one of structural misspecification with resulting bias, even introducing coefficient and equation error variances and learning (via stochastic or adaptive control) into the instrument calculation cannot unambiguously be expected to improve the feedback activism results—although a return to more normal times (with econometric forecasts as reliable as those preceding the period examined) would certainly give the advantage to feedback strategies. There was no precedent to use in evaluating the economy's response to the initial conditions and large unusual shocks of early 1974 so that policies that ignored the expected responses performed better over this period. This should not be interpreted as an indictment of active policy, but rather as a common sense warning: when predictive tools suddenly begin to miss badly, they have to be corrected (and the proper corrections were not apparent in early 1974) or the forecasting structure must be heavily discounted in policy decisions.

[34]See especially Kalchbrenner and Tinsley, p. 353.

[35]This issue was raised by the managing editor commenting on an early draft.

REFERENCES

V. Argy, "Rules, Discretion in Monetary Management and Short-Term Stability," *J. Money, Credit, Banking*, Feb. 1971, *3*, 102–22.

R. Berner et al., in *Studies in Price Stability and Economic Growth*, prepared for the Joint Economic Comm., 94th Cong., Aug. 5, 1975.

M. Bronfenbrenner, (1961a) "Statistical Tests of Rival Monetary Rules," *J. Polit. Econ.*, Feb. 1961, *69*, 1–14.

———, (1961b) "Statistical Tests of Rival

Monetary Rules: Quarterly Data Supplement," *J. Polit. Econ.*, Dec. 1961, *69*, 621-25.

G. C. Chow, "Usefulness of Imperfect Models for the Formulation of Stabilization Policies," paper presented at the Nat. Bur. Econ. Res. Conference on Stochastic Control, Palo Alto, May 1976.

C. F. Christ, "Judging the Performance of Economic Models of the U.S. Economy," *Int. Econ. Rev.*, Feb. 1975, *16*, 54-73.

J. Phillip Cooper, *Development of the Monetary Sector, Prediction and Policy Analysis in the FRB-MIT-Penn Model*, Lexington 1974.

——— **and S. Fischer**, "Stochastic Simulations of Monetary Rules in Two Macroeconometric Models," *J. Amer. Statist. Assn.*, Dec. 1972, *67*, 750-60.

——— **and C. R. Nelson**, "The *Ex Ante* Prediction Performance of the St. Louis and FRB-MIT-Penn Econometric Models and Some Results on Composite Predictors," *J. Money, Credit, Banking*, Feb. 1975, *7*, 1-31.

R. N. Cooper and R. Z. Lawrence, "The 1972-75 Commodity Boom," *Brookings Papers*, Washington, 1975, *3*, 671-723.

R. Craine, A. Havenner, and P. Tinsley, "Optimal Macroeconomic Control Policies," *Annals Econ. Soc. Measure.*, Spring 1976, *5*, 191-203.

O. Eckstein and R. Brinner, "The Inflation Process in the United States," study for the Joint Economic Comm., 92d Cong., 2d sess. 1972.

Milton Friedman, "Statement on Conduct of Monetary Policy," study for the Committee on Banking, Housing, and Urban Affairs, U.S. Senate, Nov. 6, 1975.

——— **and Walter W. Heller**, *Monetary vs. Fiscal Policy*, New York 1969.

R. J. Gordon, "Alternative Responses of Policy to External Supply Shocks," *Brookings Papers*, Washington 1975, *1*, 183-204.

E. M. Gramlich, "The Optimal Timing of Unemployment in a Recession," *Brookings Papers*, Washington 1975, *1*, 167-83.

D. Hathaway, "Food Prices and Inflation," *Brookings Papers*, Washington 1974, *1*, 63-103.

A. R. Holmes, "Monetary Policy in a Changing Financial Environment: Open Market Operations in 1974," *Fed. Res. Bull.*, Washington 1975, *61*, 197-208.

J. Kalchbrenner and P. Tinsley, "On the Use of Feedback Control in the Design of Aggregate Monetary Policy," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 349-55.

J. Kmenta and P. Smith, "Autonomous Expenditures Versus Money Supply: An Application of Dynamic Multipliers," *Rev. Econ. Statist.*, Aug. 1973, *55*, 299-307.

F. Modigliani, "Some Empirical Tests of Monetary Management and of Rules Versus Discretion," *J. Polit. Econ.*, June 1964, *72*, 211-45.

T. Muench et al., "Tests for Structural Change and Prediction Intervals for the Reduced Forms of Two Structural Models of the U.S.: The FRB-MIT and Michigan Quarterly Models," *Annals Econ. Soc. Measure.*, July 1974, *3*, 491-519.

A. Norman, "On the Relationship Between Linear Feedback Control and First Period Certainty Equivalence," *Int. Econ. Rev.*, Feb. 1974, *15*, 209-15.

A. M. Okun, "A Postmortem of the 1974 Recession," *Brookings Papers*, Washington 1975, *1*, 227-35.

———, "Fiscal Monetary Activism: Some Analytical Issues," *Brookings Papers*, Washington 1972, *1*, 123-72.

G. L. Perry, (1975a) "Policy Alternatives for 1974," *Brookings Papers*, Washington 1975, *1*, 227-35.

———, (1975b) "Understanding World Inflation," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 120-24.

J. Pierce and J. Enzler, "The Effects of External Inflationary Shocks," *Brookings Papers*, Washington 1974, *1*, 13-61.

W. Poole, "Money Growth During Business Cycle Expansions," prelim. draft prepared for the Spring 1973 Committee on Financial Analysis.

———, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quart. J. Econ.*, May 1970, *74*, 197-216.

Board of Governors of the Federal Reserve System, "Record of Policy Actions of the Federal Reserve Open Market Committee," *Fed. Res. Bull.*, Washington, various years.

———, "Numerical Specification of Financial Variables and their Role in Monetary Policy," *Fed. Res. Bull.*, Washington 1974, *60*, 333–37.

———, *Quarterly Econometric Model Data Directory*, Jan. 1975.

———, *Quarterly Econometric Model Equations*, Jan. 1975.

Business Week, "Economic Forecasts for 1974," Dec. 22, 1973.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington 1973.

Factor Abundance and Comparative Advantage

By JON HARKNESS*

Over the past twenty-five years, a great deal of effort has been devoted to expanding and testing the factor-proportions trade model and to identifying America's relatively abundant factors of production. Such efforts, of course, stem from Wassily Leontief's pioneering study which overturned the conventional wisdom by suggesting that either the Heckscher-Ohlin trade model is invalid or the United States is capital poor. However, only two firm conclusions emerge from the subsequent glut of research. First, conceptually, the multifactor-proportions trade model is capable of explaining only indirect factor trade and not commodity trade. Second, empirically, the United States appears to be a net exporter of human capital. All other results are frustratingly inconclusive.[1] Three fundamental questions remain unanswered: What, if any, is the theoretic relationship between relative factor endowments and commodity trade? Which factors are relatively abundant in the United States? Does the factor-proportions model adequately explain *U.S.* trade in commodities or (indirectly) in factors?

The inconclusiveness of the empirical studies to date results from the fact that these three questions may not be independent. For example, attempts to test the factor-proportions model, both for the United States and other nations, have generally been based on unwarranted a priori assumptions regarding relative factor abundances. On the other hand, attempts to measure relative factor abundances have been based on the equally unwarranted a priori assumption that the factor-proportions model is valid. Further, both types of study often assume that the factor-proportions model holds particular implications for the patterns of commodity and/or factor trade which it does not.

This paper attempts to answer all three questions. Section I develops the factor-proportions model for a many-goods, many-factors world in which factor prices are not necessarily equalized. It examines the theoretic relationship between both net commodity trade and net factor-service trade, on the one hand, and relative factor abundances, on the other. First, it shows, by restating the contributions of Jaroslav Vanek and Trent Bertrand, that a nation's relative factor abundances can be determined from its net flows of factor services through trade. This is conveniently referred to as the Vanek-Bertrand (V-B) Theorem and I show precisely how it will be used to estimate America's relative factor-abundance ranking. Second, by use of a linear regression equation, it shows how net commodity exports can be predicted from knowledge of the intensity of use of factors in producing such commodities. It is argued that a ranking of the coefficients on each factor intensity in the regression equation will duplicate a ranking of the corresponding relative factor abundances. This latter result is called the Heckscher-Ohlin (H-O) Theorem. Section II uses the above results to estimate America's relative abundances of a wide variety of factors and to test the H-O Theorem. Conclusions are found in Section III.

For concreteness, the model is developed in the particular context of *U.S.* trade while its somewhat restrictive assumptions are designed to conform to the data available for empirical application of the model.

*Associate professor, McMaster University. I would like to thank, without implicating, Robert Baldwin, Peter Kenen, John Kennan, Les Robb, Jim Williams, and the managing editor for many helpful comments.

[1]There are so many good reviews of this literature that another is not necessary here. For examples, see Robert Baldwin, Gary C. Hufbauer, or Robert M. Stern.

I. The Factor-Proportions Model

A. *Assumptions and Variables*

To begin our development of the factor-proportions model, consider a two-nation world with N produced commodities and M primary factors, where $N > M$. Factors are internationally immobile and homogeneous between the two nations. Let nation 1 be the United States and nation 2 be the rest of the world (ROW).

Assumption 1: *Existence of a general competitive equilibrium of the sort described by Paul Samuelson.*

Assumption 2: *Constant returns to scale in the production of all commodities in both nations.*

Assumption 3: *Identical and homothetic preferences in the two nations.*

Assumption 4: *Identical relative value shares to any given factor in the production of any given commodity in both nations.*

These last two assumptions are in general unnecessary to the theory. But without them the available data will not permit empirical implementation of the model. A more general model is developed elsewhere.[2] Assumption 3 implies that the average propensity to consume any given commodity is identical in the two nations. Assumption 4 will hold under a variety of conditions. Two of these are: (i) identical Cobb-Douglas technology in the two nations or (ii) identical technology and complete factor-price equalization in the two nations.[3] Finally, note that the usual additional assumptions necessary to generate complete factor-price equalization are not adopted. There may exist impediments to trade such as transportation costs or tariffs. Moreover, Assumption 4 does not preclude the possibility that technology is nonidentical in the two nations.

Define the following variables, where the superscript refers to nation i:

g^i_m = value of the total supply of the mth factor's services

Y^i = gross national product

y^i_n = value of output of the nth commodity

x^i_n = $X^i_n - M^i_n$ = value of net exports of the nth commodity, where X^i_n and M^i_n are, respectively, values of gross exports and imports of the nth commodity

Each of these value flows is measured in nation i's prices, except gross imports which are valued in the producing nation's prices. Gross import values therefore exclude transportation costs or tariffs. Throughout we are concerned only with the values of these (or any other) variables obtaining in the general equilibrium. Finally, we have the following two variables each of which, by Assumptions 3 and 4, have common values in the two nations:

d_n = the average propensity to consume the nth commodity, so that d_nY^i is nation i's total value of consumption of commodity n

f_{mn} = the relative value share of factor m in producing good n

Without loss of generality, it is assumed all production is completely vertically integrated so there exist no intermediate inputs.[4] Henceforth, call f_{mn} a "factor intensity" since it gives the total value of the mth factor's services used in producing a "dollar's worth" of the nth commodity.

[2] See the author where it is shown that none of the theoretic results is necessarily affected when Assumptions 3 and 4 fail to hold.

[3] It will be apparent later that for our purposes it will be sufficient that these factor shares be highly correlated between the two nations.

[4] To the extent that there do exist intermediate inputs, f_{mn} is a total (i.e., direct plus indirect) factor share. See James R. Williams for a model which explicitly considers the case of intermediate inputs.

B. *Measuring Relative Factor Abundances by the V-B Theorem*

I now briefly show how the United States' relative factor-abundance ranking can be estimated from its net flows of factor services through trade. First, I define what is meant by "relative factor abundances." Following Vanek, order relative factor endowments according to

$$(1)\quad r_m = g_m^1/g_m^2 \qquad (m = 1,\ldots,M)$$

where $r_m > r_j$ indicates the United States has an abundance of the services of factor m relative to that for factor j, vis-à-vis *ROW*. An equivalent ordering will be one based on

$$(2)\quad r_m^o = r_m - Y^1/Y^2 \qquad (m = 1,\ldots,M)$$

since $r_m^o > r_j^o$ whenever $r_m > r_j$. Moreover, when $r_m^o > 0$, the mth factor is said to be "relatively abundant" and vice versa. Henceforth, r_m^o will be called a "relative factor endowment."

Second, for nation i, define D_m^i and T_m^i as the total value of the mth factor's services embodied in, respectively, total commodity consumption and net exports. Then, since f_{mn} gives the mth factor-service content of a dollar's worth of output, consumption, or net exports of commodity n, regardless of where it is produced:

$$(3)\quad D_m^i = \sum_{n=1}^{N} f_{mn}\cdot(d_n Y^i)$$

$$(4)\quad T_m^i = \sum_{n=1}^{N} f_{mn}\cdot x_n^i$$

$$(i = 1,2);\ (m = 1,\ldots,M)$$

Now, the assumption of a general equilibrium in which all factor markets clear implies two relationships. First, the value of the United States' consumption plus net exports of any factor's services must equal the total value of her supply of such services.

$$(5)\quad g_m^1 = D_m^1 + T_m^1 \qquad (m = 1,\ldots,M)$$

Second, the *U.S.* plus *ROW* consumptions of any factor's services must sum to the total world supply of such services, $g_m^1 + g_m^2$. Then,

$$(6)\quad g_m^1 + g_m^2 = D_m^1 + D_m^2 \qquad (m = 1,\ldots,M)$$

It can now be seen that the United States will consume a fixed proportion of the total world supply of every factor's services. From expression (3), we find that $D_m^2 = Y^2D_m^1/Y^1$, since factor shares and average propensities to consume are identical in the two nations. Substitute this into expression (6) and rearrange terms:

$$(7)\quad D_m^1 = [Y^1/(Y^1 + Y^2)]\cdot(g_m^1 + g_m^2) \qquad (m = 1,\ldots,M)$$

The V-B Theorem can now be directly derived. Substitute expression (7) into equation (5) and rearrange terms:

$$(8)\quad T_m^1/g_m^1 = 1 - \alpha - \alpha g_m^2/g_m^1 = 1 - \alpha - \alpha/r_m = (1-\alpha)r_m^o/r_m \qquad (m = 1,\ldots,M)$$

where r_m^o is given by expression (2) and $0 < \alpha = Y^1/(Y^1 + Y^2) < 1$. For simplicity, define the left-hand side of this expression as T_m^o, the "proportionate net export of factor m's services" and write the right-hand side as g_m^o. Then,

$$(9)\quad T_m^o = g_m^o \qquad (m = 1,\ldots,M)$$

It can be seen that g_m^o is a monotonic transformation of r_m^o (or r_m): the sign and rank order of any g_m^o will duplicate those of r_m^o. Thus expression (9) produces the V-B Theorem: $T_m^o \gtrless T_j^o$ if and only if $r_m \gtrless r_j$ while $T_m^o \gtrless 0$ if and only if $r_m^o \gtrless 0$. In the general equilibrium, an ordering of the United States' proportionate net exports of factor services will duplicate an ordering of the corresponding relative factor endowments while her relatively (dis)abundant factors will be exported (imported).[5]

[5]Since factor prices are not necessarily equalized between the two nations, relative factor endowments are endogenous. Thus, the V-B Theorem merely characterizes the general equilibrium. Trade might result solely from technological differences with identical physical factor endowments in the two nations or from differing endowments and identical technology or from

In other words, the sign and rank order of any relative factor endowment, r_m^o, will duplicate those of the corresponding net export of factor services, T_m^o. Thus, America's relative factor-abundance ranking can be determined by computing the values of the T_m^o's. Given expression (4), these can be computed according to

$$(10) \quad T_m^o = T_m^1/g_m^1 = \sum_{n=1}^{N} f_{mn} \cdot x_n^1/g_m^1 \qquad (m = 1, \ldots, M)$$

All the data necessary to such a computation exist.

C. *Commodity Trade and the H-O Theorem*

We now investigate the relationship between commodity trade and relative factor abundances. The usual way of stating the H-O Theorem involves relative factor abundances on the one hand and relative factor intensities of production on the other. For example, in a two-good, two-factor world a nation will export (import) the good whose production is the more (less) intensive in the use of the relatively abundant factor. The difficulty encountered in attempting to extend this result to the case of many factors is well known. With more than two factors, it is not possible to provide a unique ranking of technologies based on relative factor intensities and, therefore, the commodity composition of trade can no longer be determined by reference to factor intensities alone.

Nonetheless, I propose to extend the factor-proportions model to a many-factors, many-goods world by using ordinary least squares (*OLS*) methods to estimate the following linear relationship between *U.S.* "proportionate net commodity exports," x_n^1/y_n^1, and all factor intensities:

$$(11) \quad x_n^1/y_n^1 = \beta_1 \cdot f_{1n} + \ldots + \beta_M \cdot f_{Mn} + e_n \qquad (n = 1, \ldots, N)$$

where e_n is an unspecified error. Thus, given random or "small" errors, I hypothesize that proportionate net commodity exports can be predicted from knowledge of factor intensities alone. Moreover, I assert that the sign and rank order of any *OLS*-estimated regression coefficient $\hat{\beta}_m$, will duplicate those of the corresponding relative factor endowment r_m^o; i.e., $\hat{\beta}_m \gtrless \hat{\beta}_j$ whenever $r_m^o \gtrless r_j^o$. This may be called the H-O Theorem: the "effect" of factor intensities on proportionate net commodity exports rises monotonically, across factors, with relative factor abundance, while this effect is positive (negative) for relatively abundant (scarce) factors.

I now provide a somewhat rigorous, though intuitive, justification of this model and the assertions regarding the H-O Theorem. A detailed treatment is found in my cited working paper. There are three critical features to the model. First, equation (11) has a zero intercept. This results from the assumption of constant returns to scale whereby the right-hand variables, factor intensities, sum to unity for each commodity. It is well known that *OLS* estimation of such an equation will require that the intercept be suppressed. Second, the model does not predict net commodity exports per se. Commodity trade balances have been scaled to control for the effects of differing output values. Third, equation (11) contains all factor intensities and is to be estimated using *multiple* regression methods to control for what the trade literature calls "factor complementarities." The explanation and justification of these latter two features will require some additional algebra.

By definition, the total use of factor m's services in the *U.S.* production of the nth commodity will be $f_{mn}y_n^1$. Moreover, since all factor markets clear, the use of any given

differing endowments and technology. In any case, equation (9) will hold in the general equilibrium. Thus, the V-B Theorem does not depend upon the validity of any particular trade model. Also note that it is common in the trade literature to attempt to infer relative factor endowments from T_m^1, not from T_m^o, which is, a priori, not possible.

factor in the *U.S.* production of all commodities must sum to its total *U.S.* supply. Thus,

$$(12)\quad g_m^1 = \sum_{n=1}^{N} f_{mn} \cdot y_n^1 \qquad (m = 1, \ldots, M)$$

Further, since all production is subject to constant returns to scale and there exists perfect competition, factor shares sum to unity in the production of every commodity: that is,

$$\sum_{j=1}^{M} f_{jn} = 1$$

for all n. Then, multiplying y_n^1 by a "well-chosen one," expression (12) becomes

$$(13)\qquad g_m^1 = \sum_{n=1}^{N} f_{mn} \cdot y_n^1 \Big(\sum_{j=1}^{M} f_{jn}\Big) \qquad (m = 1, \ldots, M)$$

Recalling that $T_m^o = T_m^1/g_m^1$, equation (9) which produced the V-B Theorem indicates that $T_m^1 = g_m^o \cdot g_m^1$. Substitute (13) into this expression:

$$(14)\quad T_m^1 = g_m^o \cdot \sum_{n=1}^{N} f_{mn} \cdot y_n^1 \cdot \Big(\sum_{j=1}^{M} f_{jn}\Big) \qquad (m = 1, \ldots, M)$$

Finally, setting $i = 1$, expression (4) gives an alternate definition of T_m^1:

$$(15)\quad T_m^1 = \sum_{n=1}^{N} f_{mn} \cdot x_n^1 \qquad (m = 1, \ldots, M)$$

These two alternate expressions for T_m^1 can be used to investigate the properties of the *OLS* estimate of equation (11). First, since all product markets clear, *U.S.* commodity trade will be balanced and, therefore,

$$\sum_{n=1}^{N} x_n^1 = 0$$

This implies, by expression (15), that T_m^1/N is the simple covariance across commodities between x_n^1 and f_{mn}. Since all output values and factor intensities are nonnegative, it will be apparent from (14) that the sign of this covariance is identical to that for g_m^o, and therefore to that for r_m^o. Hence, there is a positive (negative) correlation between net commodity exports and the intensity of use of (dis)abundant factors. This is a weak version of the H-O Theorem. However, this covariance (or correlation) will not necessarily be larger the larger is its corresponding relative factor endowment r_m^o. Its relative size depends among other things on the levels of the output values y_n^1, as seen from expression (14). Intuitively, the world demand and, therefore, the domestic output of goods using the most abundant factor(s) the most intensively may be relatively small. Such goods will, on average, be exported but their exports will not necessarily be the largest. It might be expected, however, that such goods will export the largest proportion of their output.

We can control for this effect by holding all output-values constant at the level, say, unity. Setting y_n^1 equal to unity for all n, and dividing by N, equation (14) yields an expression for T_m^*, the (partial) covariance between x_n^1 and f_{mn}, holding all output values constant at unity.

$$(16)\qquad T_m^* = g_m^o \cdot \sum_{j=1}^{M} \sigma_{mj}$$

where

$$(17)\quad \sigma_{mj} = \sum_{n=1}^{N} f_{mn} \cdot f_{jn}/N \quad (m,j = 1, \ldots M)$$

Essentially, all outputs have been scaled down to the unit-*value* isoquant. By the assumption of constant returns to scale (i.e., linearly homogeneous production functions), this will leave all factor intensities unaffected. Thus, in the definition of T_m^1 given by expression (15), only the x_n^1's are affected by such a scaling. Given the above discussion, we will assume *as a first approximation* that the proportion of any output value entering net trade is unaffected by this scaling. This implies that net exports of the nth commodity will be x_n^1/y_n^1 when all output values are scaled to unity.[6] Thus,

[6]For our purposes, it would be sufficient if, when y_n^1 is held constant, the value of the nth commodity's net exports were monotone in x_n^1/y_n^1. Moreover by implication all foreign output values have been similarly scaled.

from (15), an alternate expression for T_m^* will be

$$(18)\quad T_m^* = \sum_{n=1}^{N} f_{mn}\cdot(x_n^1/y_n^1)/N \qquad (m = 1,\ldots,M)$$

Now, suppose each parameter of equation (11) were estimated using *simple* least squares techniques. By well-known formulae for simple regression equations forced through the origin and by definitions (18) and (17), the simple regression coefficient on the mth factor intensity will be: $\hat{\beta}_m = T_m^*/\sigma_{mm}$. Substitute (16) into this expression and rearrange terms.

$$(19)\quad \hat{\beta}_m = g_m^o + g_m^o\cdot\Big(\sum_{j\neq m}^{M} \sigma_{mj}/\sigma_{mm}\Big) \qquad (m = 1,\ldots,M)$$

The coefficient $\hat{\beta}_m$ is a descriptive statistic summarizing the effect of the mth factor intensity on proportionate net commodity exports. If it were the case that $\hat{\beta}_m \gtreqless \hat{\beta}_j$ whenever $g_m^o \gtreqless g_j^o$ (i.e., $r_m^o \gtreqless r_j^o$), the H-O Theorem as stated above would hold on these parameter estimates. But, this cannot be the case a priori, since the relative sizes of these coefficients depend not only on the relative sizes of the g_m^o's but also on all the cross-product terms σ_{mj}. This is the problem of factor complementarities, noted above. Such complementarities arise, for example, when there is positive association across goods between the intensity of use of, say, the most and the least abundant factor. Thus, on average, goods which intensively use an abundant factor also intensively use a scarce factor and there will be two opposing effects on the level of proportionate net commodity exports. In short, with more than two factors, commodities can not be uniquely ranked according to their relative factor intensities. It can be seen that such complementarities would be absent if, for example, the terms

$$\sum_{j\neq m}^{M} \sigma_{mj}/\sigma_{mm}$$

were identical for all m. Then $\hat{\beta}_m$ would be monotone in g_m^o and the H-O Theorem would hold.

All this suggests that, from the point of view of testing the H-O Theorem, a method must be used in estimating equation (11) which controls for factor complementarities; that is, for systematic association, across goods, among factor intensities. Such a method is, of course, *multiple* regression analysis which would attempt to remove the effects of collinearity among the right-hand variables of equation (11) on the estimate of β_m. Given that factor complementarities can be controlled through multiple regression analysis, the H-O Theorem predicts that the sign and rank order of any multiple regression coefficient $\hat{\beta}_m$ will duplicate those of the corresponding relative factor endowment r_m^o.[7]

Finally, since the extent of factor complementarities is unknown a priori, a valid test of the H-O Theorem will require that all factor intensities be included in the multiple regression equation. However, we have seen that the existence of factor complementarities (or, for that matter, differing levels of domestic output values) have absolutely no bearing on the V-B Theorem. Thus, were data available only on a subset of the M factor intensities, the V-B Theorem can still be used to provide a valid partial ordering of this subset of relative factor endowments. But a valid partial ordering on the regression coefficients can not be obtained, a priori.

II. Empirical Implementation of the Model

The empirical work is now straightforward. First, compute the proportionate net export of factor services T_m^o, for every factor. By the V-B Theorem, the signs and ranks of these will duplicate those of the corresponding relative factor endowments

[7] In essence, multiple regression overcomes the problem of providing a unique ranking of technologies based on factor intensities when there are more than two factors. By holding all other factor intensities constant, it ranks goods according to the intensity of use of the mth factor relative to a constant composite of all other factor intensities.

r^o_m. Hence, these will produce an estimate of the United States' relative factor-abundance ranking. Second, using *OLS*, estimate equation (11) which, recall, alleges that proportionate net commodity exports can be "predicted" solely from knowledge of factor intensities, provided the error term is random or "small." This qualification on the errors means that the possible effects of omitted variables, if any, are commodity specific and, consequently, such variables do not vary systematically across goods with proportionate net exports. The R^2 value of the overall regression will measure the "goodness of fit" of this factor-proportions model to the *U.S.* data. Finally, the H-O Theorem is tested by comparing the signs and rank orders of the individual regression coefficients with those of the corresponding relative factor endowments, as inferred from the corresponding proportionate net export of factor services. According to our theorems, the sign and rank order of any regression coefficient $\hat{\beta}_m$ should duplicate those of the corresponding proportionate net export of factor services T^o_m.

A. *Significance Tests*

While the above model is purely descriptive, it is clear that the real world is not in fact deterministic. Randomness is introduced into the model by imposing the classical least squares assumptions on the error of equation (11). Let the error vector $e = (e_1, \ldots, e_N)$ be distributed multivariate normal with mean-vector zero and variance $\sigma^2 \cdot I$, where I is the $N \times N$ identity matrix. The usual significance tests can now be performed on the *OLS* estimate of equation (11).

Moreover, this will now permit significance tests on the relative factor endowments, as inferred from T^o_m. Since x^1_n/y^1_n has a random component, in accord with the error in equation (11), T^o_m as defined by (10) is likewise random. Thus, the observation that $T^o_m \neq T^o_j$ and, therefore, the inference that $r^o_m \neq r^o_j$ may be due solely to chance. Given the distribution of the T^o_m's, tests for significant differences among the inferred relative factor endowments can be constructed.

Let $\hat{x}^1_n = x^1_n - \hat{e}_n \cdot y^1_n$ be the value of x^1_n predicted from *OLS* estimation of equation (11), where $\hat{e}_n$ is the calculated regression error. Then by expression (10),

$$(20) \quad T^o_m = \sum_{n=1}^{N} f_{mn}(\hat{x}^1_n + \hat{e}_n \cdot y^1_n)/g^1_m \qquad (m = 1, \ldots, M)$$

Define $T^o = (T^o_1, \ldots, T^o_M)$. Given the assumed distribution of e and assuming g^1_m is fixed in repeated trials,[8] it is readily confirmed that T^o is distributed multivariate normal with mean vector $ET^o = (ET^o_1, \ldots, ET^o_m)$ and variance $\sigma^2 VV'$, where V is the $M \times N$ matrix with typical element $f_{mn}y^1_n/g^1_m$ and

$$(21) \quad ET^o_m = \sum_{n=1}^{N} f_{mn}\hat{x}^1_n/g^1_m \quad (m = 1, \ldots, M)$$

Significance tests on the relative factor endowments as inferred from ET^o_m can now be constructed in accord with methods described by Donald Morrison. Henceforth, refer to T^o_m as the "observed" and to ET^o_m as the "estimated" proportionate net export of the mth factor's services. Note that to infer relative factor abundances from ET^o_m rather than from T^o_m is conditional on the validity of the factor-proportions model, for it requires that the errors of equation (11) be randomly and independently distributed with zero mean. This, recall, is identical to assuming that only factor intensities have a systematic effect on proportionate net commodity exports.

B. *Empirical Results for the United States*

The study was conducted for the year 1958 on all *U.S.* commodities aggregated according to the industry classification of the 1958 *U.S.* Input-Output (*I-O*) Table. This choice of sample year is dictated by

[8]These assumptions also imply that y^1_n and f_{mn} are nonrandom. Since e_n is normally distributed, so is x^1_n/y^1_n. This can be true only if y_n is nonrandom which implies by relation (12) that f_{mn} is nonrandom.

TABLE 1—OBSERVED RELATIVE FACTOR ABUNDANCES FOR THE UNITED STATES, 1958, AND TESTS OF THE H-O THEOREM

Factor	1 T_m^o	2[a] $\rho(\beta_m^1)$	3[a] $\rho(\beta_m^2)$	4[a] $\rho(\beta_m^3)$
1. Coal	.340	5	5	2
2. Chemical and Fertilizer Minerals	.185	6	7	7
3. Scientists and Engineers	.079	2	1	1
4. Cropland	.068	9	8	4
5. Inventories	.067	-	11	-
6. Skilled Craftsmen	.051	3	4	5
7. Total Capital	.049	-	-	11
8. Physical Capital	.036	1	2	-
9. Operatives	.013	8	9	10
10. Clerks and Salesmen	.009	4	3	3
11. Farmers and Farm Labor	.008	-	-	-
12. Nontechnical Professionals and Managers	.006	14	16	15
13. Unskilled Labor	.003	12	12	12
14. Petroleum and Natural Gas	.000	11	13	6
15. Pastureland	-.002	10	10	9
16. Stone and Clay	-.273	13	14	13
17. Forests	-.360	7	6	8
18. Iron and Ferro-Alloy Ores	-.508	16	17	16
19. Nonferrous Metal Ores	-.571	15	15	14
Spearman: SR[b]		.72[c] (2.8)	.65[c] (2.6)	.70[c] (2.7)
Kendall: KR[b]		.50[c] (2.7)	.46[c] (2.6)	.52[c] (2.8)

[a] $\rho(\beta_m^i)$ gives the rank, in descending order, of the corresponding parameter estimates of regression equation i, as recorded in Table 2; $i = 1, 2, 3$.
[b] The corresponding z-statistic for Spearman's (SR) and Kendall's (KR) rank correlation coefficients are recorded in parentheses.
[c] Significant at the 99 percent confidence level.

the availability of all necessary data. Data sources are in the Appendix.

The eighteen mutually exclusive and (almost) exhaustive factors considered are listed in the first column of Table 1. The various natural resources are primary inputs, defined as subsoil productive assets.[9] Physical capital is defined as physical plant and equipment. Initially, it is assumed that "inventories" are not a factor, since they are unlikely directly to enter a neoclassical production function. Later, they are a factor and are imperfectly substitutable for physical capital, or perfectly substitutable, for which we have the variable "total capital" being the sum of physical capital plus inventories. Factor intensities are measured by the total (i.e., direct plus indirect) factor rewards per dollar delivery to final demand, computed in the usual fashion using the Leontief inverse of the I-O Table. These would be relative factor shares were each industry completely vertically integrated. In accord with the theory, all magnitudes entering trade are valued in the prices prevailing in the country of origin.

The observed proportionate net exports of factor services T_m^o, as defined by equation (10), are recorded in rank order in column 1 of Table 1. According to the V-B Theorem, the signs and ranks of these will duplicate those of the relative factor endowments r_m^o. We have then the (almost) complete relative factor-abundance ranking for the United States in 1958. Factors for which

[9] These are pristine resources, not resource products. See the Appendix for the definition of resource inputs.

TABLE 2—REGRESSION EQUATIONS RELATING *U.S.* PROPORTIONATE NET EXPORTS BY INDUSTRY TO FACTOR INTENSITIES, 1958[a]

Factor Intensity	Equation 1	Equation 2	Equation 3
Physical Capital	10.208[c]	10.248[c]	-
	(2.82)	(2.93)	
Inventories	-	-1.313[c]	-
		(2.27)	
Total Capital	-	-	-1.172
			(1.20)
Pastureland	-4.505	-1.261	-.055
	(1.47)	(0.58)	(0.03)
Cropland	-0.901	0.978	4.699
	(0.57)	(0.03)	(1.60)
Forests	1.095	1.998	0.428
	(0.47)	(1.61)	(0.46)
Iron and Ferro-Alloy Ores	-105.921[c]	-105.100[c]	-96 670[c]
	(22.36)	(23.19)	(23.12)
Nonferrous Metal Ores	-7.315[c]	-7.090[c]	-5.514[c]
	(7.42)	(7.77)	(6.51)
Coal	2.294	2.781	7.158
	(0.39)	(0.51)	(0.92)
Petroleum and Natural Gas	-4.586[c]	-4.508[c]	1.289
	(2.43)	(2.33)	(0.92)
Stone and Clay	-4.982[c]	-4.815[c]	-4.794[c]
	(8.83)	(9.23)	(7.66)
Fertilizer and Chemical Minerals	1.513	1.802[b]	1.041[b]
	(1.47)	(1.82)	(1.67)
Scientists and Engineers	8.087[c]	16.245[c]	8.642[b]
	(2.05)	(3.13)	(1.71)
Non-Technical Professionals and Managers	-6.048	-13.008[c]	-8.295
	(1.60)	(2.73)	(1.54)
Clerical and Sales Workers	3.860	8.842[b]	6.803
	(0.86)	(1.98)	(1.47)
Skilled Craftsmen	4.672[c]	5.902[c]	3.932[c]
	(2.66)	(3.37)	(2.35)
Operatives	0.465	0 598	-0.767
	(0.42)	(0.47)	(1.32)
Unskilled Labor and Service Workers	-3.889[c]	-2.553	-3.071
	(2.09)	(1.33)	(1.08)
Constant	-0.731[b]	-0.883[c]	-0.084
	(1.99)	(2.47)	(0.27)
R^2	0.948[c]	0.946[c]	0.940[c]
F	69.711	66.813	55.384

[a]The dependent variable "proportionate net commodity exports" for any given industry is measured by x_n/y_n, net exports as a proportion of output.

[b]Indicates significance at the 95 percent confidence levels for the t-values (shown in parentheses) of the regression coefficients and the F-ratio on the R^2 value.

[c]Indicates significance at the 99 percent confidence levels for the t-values (shown in parentheses) of the regression coefficients and the F-ratio of the R^2 value.

T^o_m is positive are called relatively abundant and vice versa.

Table 2 presents the *OLS* estimate of equation (11). The dependent variable x_n/y_n is the value of net exports per dollar of *final* output of commodity n, which would be the total output of the industry were it completely vertically integrated. The three equations reported correspond to the three different treatments of capital noted above. It was noted earlier that from the theory, equation (11) has a zero intercept. This re-

TABLE 3—ESTIMATED RELATIVE FACTOR ABUNDANCES FOR THE UNITED STATES, 1958, AND TESTS OF THE H-O THEOREM
(Based on Regression Equation 1, Table 2)

Factor	1 ET^{o}_{m}	2 $\geq$	3 $\rho(\beta_m)$[a]	4 $sgn\ \beta_m$
Relatively Abundant				
1. Coal	.3416[c]	>	5	+
2. Scientists and Engineers	.1622[c]	>	2[d]	+
3. Chemical and Fertilizer Minerals	.1225[c]	>	6	+
4. Skilled Craftsmen	.0669[c]	>	3[d]	+
5. Physical Capital	.0377[c]	=	1[d]	+
6. Cropland	.0283[c]	>	9	−
7. Clerks and Salesmen	.0092[c]	>	4	+
8 Nontechnical Professionals and Managers	0061[c]	>	14	−
Indeterminately Abundant				
9. Operatives	.0009	=	8	+
10. Unskilled Labor	0002	=	12[d]	−
11. Petroleum and Natural Gas	0000	=	11[d]	−
12 Farms and Farm Labor	−.0118	>		
Relatively Scarce				
13 Pastureland	−.0504[c]	>	10	−
14. Stone and Clay	−.2712[c]	>	13[d]	−
15. Forests	−.3948[c]	>	7	+
16. Nonferrous Metal Ores	−.4833[c]	>	15[d]	−
17 Iron and Ferro-Alloy Ores	−.5634[c]		16[d]	−
Spearman: SR[b]	.963[c]		.774[c]	
	(3 8)		(3.0)	
Kendall. KR[b]	.851[c]		.583[c]	
	(4.7)		(3.2)	

[a] $\rho(\beta_m)$ gives the rank, in descending order, of the regression parameters from equation 1, reported in Table 2.

[b] The z-statistic for Spearman's (SR) and Kendall's (KR) rank correlation coefficients are recorded in parentheses.

[c] Indicates significance at the 99 percent confidence level.

[d] Indicates the corresponding regression parameter is significant.

sulted from the assumption of constant returns to scale whereby factor shares (i.e., the right-hand variables) sum to unity for every commodity. However, in practice, our list of factors is not exhaustive, factor shares do not in the data sum to unity, and severe collinearity problems require omission of the factor intensity "Farmers, Farm Labor and Farm Managers." These facts permit and demand the estimation of equation (11) with an intercept. From the point of view of the factor-proportions model, the regression constant will simply be the implied regression coefficient on a composite factor intensity defined as the value share of all omitted factors.[10] In practice, the intercept measures the average joint effect of all omitted variables which may in fact impinge on proportionate net exports. These omissions may include variables deriving

[10] This is a well-known property of the intercept when all omitted and included variables sum to unity. Two possibly omitted factors are marine resources and urban-industrial land. Conceptually, no factor should be omitted from the regression. But "Farmers et al." is so highly complementary with "Agricultural Lands" that its separate influence can not be identified.

TABLE 4—F-STATISTICS FOR DIFFERENCES BETWEEN PROPORTIONATE NET FACTOR-SERVICE EXPORTS[a]

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	54.7																
2	27.8	49.2															
3	32.7	10.9	32.4														
4	43.8	46.3	18.1	19.4													
5	44.4	34.4	23.6	14.6	8.0												
6	44.6	18.5	13.7	4.9	1.0	5.7											
7	46.0	42.2	24.5	24.6	4.5	3.6	10.4										
8	45.5	43.5	25.5	28.0	4.9	2.4	7.8	10.1									
9	51.4	49.8	33.5	33.0	7.0	12.3	21.6	19.4	0.5								
10	30.6	47.9	35.1	44.5	9.9	30.7	61.4	60.8	0.3	0.3							
11	54.0	47.5	31.4	18.5	5.4	7.6	9.5	9.7	0.0	0.1	0.9						
12	37.9	23.7	19.7	13.2	25.6	10.9	10.5	10.1	1.2	1.7	1.1	1.0					
13	39.6	27.1	23.8	19.6	25.5	19.3	18.4	18.0	7.8	10.1	13.0	25.4	7.6				
14	95.4	146.8	142.0	162.1	60.6	130.8	132.7	138.3	77.7	86.9	106.3	44.7	25.5	78.7			
15	96.4	107.5	102.7	112.6	76.1	104.9	114.6	113.8	70.6	101.4	107.6	64.2	43.9	26.5	71.0		
16	123.5	196.2	141.7	154.3	71.3	108.2	124.9	126.3	143.8	112.6	106.7	58.4	42.0	54.1	14.9	147.7	
17	143.4	267.8	181.7	255.7	90.3	153.9	186.0	189.6	174.6	158.3	159.6	75.0	54.4	96.0	31.3	24.5	179.8

[a]The row and column numbers refer to factors as numbered in Table 3. The 5 and 1 percent confidence limits are 1.88 and 2.44, respectively.

from competing theories of comparative advantage.[11]

At present, no interest attaches to the individual parameters of the regression equations. As a test of the factor-proportions model, note that factor intensities as a group "explain" about 95 percent of the variation in proportionate net commodity exports across U.S. industries as indicated by the (significant) R^2 values for the overall regression. The regression constant is significantly nonzero in two of the three equations. It is not possible, however, to disentangle the contribution to the constant of omitted factor intensities from that, if any, of omitted variables deriving from other possible models. Nonetheless, the U.S. data are consistent with the factor-proportions trade model.

Given this result, it is now legitimate to estimate the relative factor-abundance ranking by computing the magnitudes ET^o_m, as defined by expression (21). Using regression equation 1 from Table 2 to estimate net commodity exports, the computed values of ET^o_m are tabulated in rank order in column 1 of Table 3. It is of interest to know how well the relative factor-abundance ranking inferred from these estimates "fits" that inferred from the "observed" values T^o_m, recorded in Table 1. Spearman's and Kendall's rank correlation coefficients, SR and KR,[12] between these two alternate rankings are recorded at the bottom of column 1 of Table 3. They indicate a good fit of the estimated to the observed ranking.[13]

Using the assumptions regarding the underlying distribution of the vector T^o, diagonal and off-diagonal entries in Table 4 give the F-statistic for the null hypothesis that, respectively, $ET^o_m = 0$ and $ET^o_m - ET^o_j = 0$. A majority of the ET^o_m's are significantly different from zero and from each other at high confidence levels. But from the V-B Theorem, this implies a majority of the relative factor endowments r^o_m, are significantly different from zero and from each other.

[11]For a summary of competing hypotheses see Baldwin, Hufbauer, or Stern.

[12]Discussion of these two rank correlation coefficients is to be found in Frederick Mills.

[13]While Spearman's is the better-known coefficient to economists, its z-statistic is unreliable for sample sizes less than twenty.

The results of Table 4 are summarized in Table 3. Footnote key c on the magnitude ET^o_m in column 1 indicates it is significantly nonzero. An inequality (equality) sign attached to a factor in column 2 indiciates it is (in)significantly more abundant than the following factor. The stub headings in the table divide factors into three groups where ET^o_m is significantly greater than, equal to, or less than zero corresponding to factors which are, respectively, relatively abundant, indeterminately abundant, and relatively scarce. We now have a statistical estimate of the (almost) complete relative factor abundance ranking for the United States in 1958.

The H-O Theorem can now be tested. The theorem's first prediction is that a ranking of the regression coefficients will duplicate a ranking of the corresponding relative factor endowments as inferred from either T^o_m or ET^o_m. The ranks of the corresponding parameters of regression equation 1 are recorded in column 2 of Table 1 and column 3 of Table 3. The regression-parameter ranking does not duplicate either of the two inferred relative factor-abundance rankings. However, Spearman's or Kendall's rank correlation coefficient between the regression parameter and factor-abundance rankings, recorded as *SR* and *KR* at the bottom of column 2 of Table 1 and column 3 of Table 3, suggests that observed rank differences are atrributable to chance. Whether relative factor abundances are inferred from T^o_m or from ET^o_m, the rank correlation between the regression-parameter ranking and the relative factor-abundance ranking is significantly different from zero at a high confidence level.

Second, the H-O Theorem predicts that the signs of the regression coefficients will duplicate those of the corresponding relative factor endowments. The signs of the coefficients of regression equation 1 are recorded for each factor in column 4 of Table 3. Obviously, these do not duplicate the signs on the corresponding inferred relative factor endowments. However, for each factor where a sign difference occurs either the regression coefficient or the relative factor endowment is not significantly different from zero, in which case either one or the other can not be signed, statistically, at a generally accepted confidence level. One may infer, therefore, that observed sign differences are attributable to chance. In short, the data would appear to be consistent with the H-O Theorem.

The above calculations and tests were also performed using regression equations 2 and 3 in Table 2. These results are summarized in, respectively, Tables 5 and 6, as well as in columns 3 and 4 of Table 1. None of these results differs qualitatively from those already reported. The estimated relative factor-abundance rankings are quite similar while the H-O Theorem holds, in each case.

There is one interesting individual disparity, however, which relates to the treatment of inventories. Consider, first, the results of regression equation 2 reported in Table 5. Inventories are the fourth most abundant factor while inventory intensity has a significant negative effect on proportionate net exports. Suppose, however, that inventories are not a factor of production, but rather are held for transactions purposes so as to moderate the effects of uncertainty with respect to product demand or input supplies. It may be that industries facing greater uncertainty, *ceteris paribus*, hold larger inventories while relative uncertainty affects an industry's comparative advantage. Then, inventory intensity may be acting as a proxy for relative uncertainty which empirically has a negative impact on proportionate net exports. This may be an additional noncompeting determinant of an industry's comparative advantage not accounted for in the deterministic factor-proportions model. Of course, inventories may also be directly productive. Then, the effect of inventory intensity on proportionate net exports is a priori unknown: it will be positive on account of relative abundance but negative on account of relative uncertainty. A posteriori, relative uncertainty would appear to dominate.

TABLE 5—ESTIMATED RELATIVE FACTOR ABUNDANCES FOR THE UNITED STATES, 1958, AND TESTS OF THE H-O THEOREM
(Based on Regression Equation 2, Table 2)

Factor	1 ET^o_m	2 $\gtrless$	3 $\rho(\beta_m)$[a]	4 $sgn\ \beta_m$
Relatively Abundant				
1. Coal	.328[b]	>	5[d]	+
2. Chemical and Fertilizer Minerals	.113[b]	>	7	+
3. Scientists and Engineers	.095[b]	>	1[d]	+
4. Inventories	.048[b]	>	11[d]	−
5. Skilled Craftsmen	.027[b]	>	4[d]	+
6. Cropland	.014[c]	>	8	+
7. Physical Capital	.007[c]	>	2[d]	+
Indeterminately Abundant				
8. Non-Technical Professionals and Managers	.006	=	16[d]	−
9. Clerks and Salesmen	.003	=	3[d]	+
10. Petroleum and Natural Gas	.001	>	13[d]	−
Relatively Scarce				
11. Unskilled Labor	−.029[b]	>	12	−
12. Operatives	−.036[b]	>	9	+
13. Farmers and Farm Labor	−.042[b]	>		
14. Pastureland	−.118[b]	>	10	−
15. Stone and Clay	−.297[b]	>	14[d]	−
16. Forests	−.423[b]	>	6	+
17. Nonferrous Metal Ores	−.507[b]	>	15[d]	−
18. Iron and Ferro-Alloy Ores	−.574[b]	>	17[d]	−
Spearman. SR	.947[b]		.591[b]	
	(3.9)		(2.4)	
Kendall: KR	.856[b]		.455[b]	
	(4.9)		(2.5)	

[a] $\rho(\beta_m)$ gives the rank of the regression parameter estimates of equation 2, as reported in Table 2.
[b] Indicates significance at the 99 percent confidence level.
[c] Indicates significance at the 95 percent confidence level.
[d] Indicates the corresponding regression parameter is significant at (at least) the 95 percent confidence level.

Now, consider the results of regression equation 3 reported in Table 6, where inventories are assumed to be a factor which is perfectly sustitutable for physical capital. A version of Leontief's Paradox is observed: total capital is the fourth most abundant factor while total capital intensity has a large negative effect on proportionate net exports. This may be due to aggregation bias. A factor (physical capital) has been aggregated with a nonfactor or with an imperfectly substitutable factor (inventories). The negative effect of inventory intensity observed in regression equation 2 dominates this parameter estimate. Thus this paradox presumably results from the same considerations as those given for the case of inventory intensity alone.

For comparison with more aggregative studies, T^o_m and ET^o_m have been computed for four broadly defined factors: capital, human capital, agricultural land, and natural resources. These are tabulated in Table 7, where the three separate estimates of ET^o_m derive from the three separate regression equations. First, when capital includes inventories, all inferred relative factor-abundance rankings conform to the

TABLE 6—ESTIMATED RELATIVE FACTOR ABUNDANCES FOR THE UNITED STATES, 1958, AND TESTS OF THE H-O THEOREM
(Based on Regression Equation 3, Table 2)

Factor	1 ET^o_m	2 $\gtreqless$	3 $\rho(\beta_m)^a$	4 $sgn\ \beta_m$
Relatively Abundant				
1. Coal	.339[b]	>	2	+
2. Scientists and Engineers	.136[b]	>	1[c]	+
3. Chemical and Fertilizer Minerals	.108[b]	>	7[c]	+
4. Total Capital	.066[b]	>	11	−
5. Skilled Craftsmen	.056[b]	>	5[c]	+
6. Cropland	.053[b]	>	4	+
7. Nontechnical Professionals and Managers	.035[b]	=	15	−
8. Clerks and Salesmen	.034[b]	>	3	+
Indeterminantly Abundant				
9. Farmers and Farm Labor	.005	=		
10. Petroleum and Natural Gas	.000	=	6	+
11. Unskilled Labor	−.003	>	12	−
Relatively Scarce				
12. Operatives	−.013[b]	>	10	−
13. Pastureland	−0.59[b]	>	9	−
14. Stone and Clay	−.293[b]	>	13[c]	−
15. Forests	−.400[b]	>	8	+
16. Nonferrous Metal Ores	−.492[b]	>	14[c]	−
17. Iron and Ferro-Alloy Ores	−.584[b]		16[c]	−
Spearman: SR	.939[b]		.676[b]	
	(3.7)		(2.6)	
Kendall: KR	.772[b]		.500[b]	
	(4.3)		(2.7)	

[a] $\rho(\beta_m)$ gives the rank of the regression parameter estimates of equation 3, as reported in Table 2.
[b] Indicates significance at the 99 percent confidence level.
[c] Indicates the corresponding regression parameter is significant at (at least) the 95 percent confidence level.

conventional wisdom. Second, defining capital to exclude inventories, all but one estimate suggests human capital is relatively more abundant than physical capital. This likewise is consistent with the conventional wisdom. Third, by appropriately aggregating coefficients from the regression equation, the implied regression parameters on these four broad factor intensities can in principle be computed. In practice, the omission of the factor intensity "Farmers, Farm Labor and Farm Managers" from the regression equations prevents a fully appropriate estimate of the coefficient on human capital. Nonetheless, these aggregate parameters are recorded as β_m in Table 7. For regression equations 1 and 2, the H-O Theorem holds. It does not hold for equation 3, presumably for reasons connected with the inclusion of inventories in the definition of capital, discussed above. Finally, the regression equations could be reestimated using these four aggregate factor intensities as right-hand variables: that is, by aggregating the factors themselves rather than their regression coefficients. This has been done with the result that none of these four broad factor intensities, either separately or jointly, contributes significantly to the variation of propor-

TABLE 7—AGGREGATE RELATIVE FACTOR ABUNDANCES FOR THE UNITED STATES, 1958, AND TESTS OF THE H-O THEOREM[a]

	1	2 Regression Equation 1		3 Regression Equation 2		4 Regression Equation 3	
	T_m^o	ET_m^o	β_m	ET_m^o	β_m	ET_m^o	β_m
1. Total Capital	.049	.101[c]	-	.025[c]	8.93	.066[c]	-1.17
2. Human Capital[b]	.046	.043[c]	7.14[c]	.004[c]	16.04[c]	.034[c]	7.24[c]
3. Physical Capital	.036	.038[c]	10.21[c]	.007[c]	10.24[c]	.020[c]	-
4. Agricultural Land	.035	-.014[c]	-5.41	-.049[c]	-0.28[c]	-.041[c]	-4.64
5. Natural Resources	-.056	-.026[c]	-117.90[c]	-0.50[c]	-118.98[c]	-.094[c]	-97.06[c]

[a]Column 1 gives the "observed" factor abundances T_m^o, while columns 2-4 give "estimated" abundances ET_m^o, and aggregate regression parameters based on, respectively, regression equations 1-3, from Table 2

[b]Human Capital is defined as the sum of rewards to all labor. In computing the implied aggregate regression parameter, Human Capital excludes "Farmers, Farm Labor and Farm Managers."

[c]Indicates the accompanying magnitude is nonzero at (at least) the 99 percent confidence level.

tionate net exports across *U.S.* industries. This again is a case of aggregation bias. Clearly imperfectly substitutable factors have been aggregated which presumably has the effect of neglecting many individual factor complementarities which we have seen to be important in determining net commodity exports. In short, various types of capital, natural resources, agricultural lands, and labor skills are each sufficiently nonsubstitutable in production that none can be treated as homogenous for purposes of predicting *U.S.* commodity trade.

III. Conclusions

This paper has been concerned with two related empirical exercises. First, by applying the V-B Theorem on the factor-service content of trade, it has estimated America's relative endowments of a wide variety of factors. As was shown, this method is not conditional on the validity of any particular trade model. At a highly aggregate level, the conventional wisdom regarding America's relative factor-abundance ranking is supported. However, these aggregate results hide much. As one might have suspected, the United States is not relatively rich in all types of human capital nor is it relatively poor in all types of natural resources. Second, it has provided a version of the factor proportions model of commodity trade and its accompanying H-O Theorem. On the basis of several tests, this model is consistent with the *U.S.* data.

My results, however, do not preclude the possible coexistence of other, perhaps dynamic, determinants of *U.S.* commodity trade. However, the regression equations do suggest that the effects of other determinants, if any, are either commodity specific, having little systematic impact on proportionate net exports, across all *U.S.* industries; or captured by the coefficient on one of the factor intensities. The latter would be the case when a possibly omitted determinant is highly collinear with one of the factor intensities. For example, it is often argued that *U.S.* comparative advantage lies in the production of goods with a high research and development (*R&D*) content. Such goods may possess a temporary monopoly over "new" goods or a temporary technological advantage in the production of "old" goods. But the effect of *R&D* may be captured by the regression coefficient on "Scientists and Engineers," for it is unknown if their contribution to proportionate net exports stems from their *R&D* activity or from their direct production activity. Nonetheless, their observed relative abundance is wholly consistent with the view that the United

States has a comparative advantage in the production of "new knowledge" which is embodied in the goods she exports.[14]

Finally, at least one caveat must be explicitly attached to this study. Among other sources, aggregation bias may exist in our results due to the use of the Input-Output industry classification. Given data limitations, tests for such bias are not possible.[15]

APPENDIX

All data were aggregated and classified according to the seventy-nine industries of the 1958 *U.S.* Input-Output Table. Having computed total factor intensities, non-traded services were dropped, leaving a sample of sixty "commodities."

A. Trade Data: Commodity exports and imports are from *U.S. Exports and Imports as Related to Output, 1958.*

B. Land: Cropland and pastureland data are from Raymond Goldsmith and *Census of Agriculture, 1959.* These were converted to value shares using my estimates of 9 and 6 percent rates of return to, respectively, Cropland and Pastureland. Land is a direct input only into agriculture.

C. Capital: Data for manufactures are from *Census of Manufactures, 1958.* These are net book values of physical plant and equipment and inventories of work in progress, raw materials and finished goods. Agricultural capital estimates are from Goldsmith and *Census of Agriculture, 1959.* Sources for extractive industries include Daniel Creamer et al.; Goldsmith; *Minerals Yearbook, 1958*; *Historical Statistics of the United States.* Capital coefficients for all other industries come from Baldwin. These data were converted to value shares by assuming an 8 percent rate of return to capital, which is consistent with my own as well as other estimates in the literature.

D. Natural Resources: Forest-resource inputs come from Goldsmith. For subsoil assets, data were obtained on value-added and payments to capital and labor from sources given above. Subtracting payments to labor and capital from value-added, the residual must be the returns to the subsoil asset extracted, given constant returns to scale and perfect competition. Any subsoil asset is a direct input only into the industry which extracts it.

E. Labor Skills coefficients were kindly provided by Baldwin. These were converted to value shares using average earnings data in *Census of Population, 1960.*

[14]See Raymond Vernon or Donald Keesing for variants of the *R & D* hypothesis.

[15]Consider the *I-O* industry "Forestry and Fisheries." No data exist separately on these two dissimilar industries nor on marine-resource inputs. Thus, computation of total factor intensities using the inverse of the *I-O* Table may measure as forest-intensive industries, which are in fact marine resource intensive. This may account for Forests yielding such anomalous results in the sign and rank tests of the H-O Theorem. Further, given that inventories are held to insure certainty of supply of intermediate inputs, their levels will be highly sensitive to the actual *I-O* structure of the economy. This may bias the results pertaining to inventory intensity which is measured in accord with the theory as if all firms were completely vertically integrated.

REFERENCES

R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, *61*, 131-46.

T. J. Bertrand, "An Extension of the N-Factor Case of the Factor-Proportions Theory," *Kyklos*, July 1972, *25*, 592-96.

Daniel B. Creamer et al., *Capital in Manufacturing and Mining*, Princeton 1960.

Raymond Goldsmith, *The National Wealth of the United States in the Post-War Period*, Princeton 1962.

J. P. Harkness, "Factor Abundance and United States Comparative Advantage: A Theoretic and Empirical Study," work. paper no. 77-09, McMaster Univ. 1977.

G. C. Hufbauer, "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods," in Raymond Vernon, ed., *The Technology Factor in International Trade*, New York 1970,

145-231.

D. B. Keesing, "The Impact of Research and Development on United States Trade," *J. Polit. Econ.*, Feb. 1967, *75*, 38-48.

W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-Examined," *Proc. Amer. Phil. Soc.*, Sept. 1953, *97*, 332-49.

Frederick Mills, *Statistical Methods*, New New York 1955.

Donald F. Morrison, *Multivariate Statistical Methods*, New York 1967.

P. Samuelson, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, Feb. 1953, *21*, 1-20.

R. M. Stern, "Testing Trade Theories," work. paper no. 48, Res. Seminar Int. Trade, Univ. Michigan 1973.

George J. Stigler, *Capital and Rates of Return in Manufacturing Industries*, Princeton 1963.

William P. Travis, *The Theory of Trade and Protection*, Cambridge, Mass. 1964.

J. Vanek, "The Factor Proportions Theory: The N-Factor Case," *Kyklos*, Oct. 1968, *21*, 749-56.

R. Vernon, "International Investment and International Trade in the Product Cycle," *Quart. J. Econ.*, May 1966, *80*, 190-207.

J. R. Williams, "The Factor Proportions Theorem: The Case of M Commodities and N Factors," *Can. J. Econ.*, May 1977, *10*, 282-88.

U.S. Bureau of the Census, *Census of Manufactures*, Washington 1963.

———, *Census of Agriculture, 1959*, Washington 1963.

———, *Census of Population, 1960*, Washington 1963.

———, *U.S. Commodity Exports and Imports as Related to Output, 1958*, Washington 1962.

———, *Historical Statistics of the United States* and *Supplement*, Washington 1960.

U.S. Bureau of Mines, *Mineral Yearbook*, Washington 1959.

U.S. Office of Business Economics, "The Transactions Table of the 1958 U.S. Input-Output Study and Revised Direct and Total Requirements Data," *Surv. Curr. Bus.*, Sept. 1968, *44*, 45-49.

Anticipated Inflation and Interest Rates: Further Interpretation of Findings on the Fisher Equation

By MAURICE D. LEVI AND JOHN H. MAKIN*

This paper extends the approach first taken by Robert Mundell of employing a general equilibrium model to question the Fisher hypothesis that the real rate of interest is invariant with respect to changes in anticipated inflation. The extension involves the inclusion of a labor sector which rationalizes short-run Phillips curves of positive or negative slope and the incorporation of the effect of taxes on nominal interest as discussed by Michael Darby.

A labor market is introduced which clears when money wages demanded by labor suppliers are equal to money wages offered by labor demanders. This implies the existence of income and employment effects arising from changes in anticipated inflation whenever the elasticity of money wages demanded by labor suppliers with respect to actual inflation differs from unity.[1] For values of such an elasticity below unity, the implied positive impact upon real income arising from a rise in anticipated inflation adds to the impact upon saving arising from real balance effects and thereby increases the effect of a change in anticipated inflation upon the real rate of interest. Taxes on nominal interest earnings produce an additional effect. Our aim is to present a framework within which to consider all of these effects simultaneously.[2]

The assumption that money wages are rigid downward along with employment of the Correspondence Principle allows us to place limits upon the range of values of the elasticity of money wage demands with respect to inflation. Allowing this parameter to vary over such a range produces, along with other parameter values, comparative static results which are consistent with observable magnitudes. In particular, results are obtained describing the impact of a change in expected inflation on nominal interest and an implied impact on real interest which are consistent with the measured impact discovered by empirical researchers. However, contrary to the assumption of empirical investigators (including William Gibson; Eugene Fama; Kajal Lahiri; and John Carr, James Pesando, and Lawrence B. Smith), our results are also consistent with the hypothesis that neither the real rate of interest nor the after-tax real rate (discussed below), is independent of the expected rate of inflation.

The paper consists of three sections. Section I briefly reviews the theoretical literature questioning the independence of the real rate from the level of anticipated inflation implied by the Fisher hypothesis. We present our model and describe its implications for the relationship between nominal and real after-tax interest rates and

*Associate professor, faculty of commerce, University of British Columbia, and professor, department of economics, University of Washington, respectively. The U.S. Treasury Department and the Federal Reserve Bank of San Francisco provided financial support. We owe thanks for comments and stimulation to Charles R. Nelson, John Murray, A. L. Annanthanarayanan, and to the managing editor of this *Review*. All views expressed and any remaining errors are ours alone.

[1]The key role played in the massive contemporary literature on the Phillips curve by the elasticity of money wages demanded with respect to prices and/or anticipated prices has been ably surveyed by John Rutledge.

[2]We note that our model is consistent with a changing relationship between unemployment and inflation hypothesized by Irving Fisher (1926). This same model also produces results consistent with Fisher's *findings* on the impact of anticipated inflation upon nominal interest rates.

anticipated inflation. Section II discusses the implications of results obtained with our model for the interpretation of the estimates obtained in a number of empirical investigations of the Fisher hypothesis. Section III presents some concluding remarks.

I. A General Equilibrium Framework for Investigating the Effects of Anticipated Inflation

This investigation is by no means the first attempt to find a theoretical rationale to reconcile Fisher's hypothesis with empirical findings. Mundell showed, in a full employment world with real balance effects, that the real interest rate is dependent upon the level of anticipated inflation. Implicit in Mundell's model (since his *LM* schedule was not vertical) was a nonzero interest elasticity of money demand, a condition necessary for the link between real interest and anticipated inflation within his formulation. This requirement was made explicit by Thomas Sargent who used a model not requiring full employment, where all behavior is dependent upon distributed-lag representations of variables affecting the money and commodity markets. Sargent showed that the initial impact of a change in anticipated inflation on nominal interest is less than unity, but the full impact over long periods of time is equal to unity. Under his formulation this means that changes in anticipated inflation have only a transitory effect on the real interest rate. Ignazio Visco, with a model identical to Sargent's save for the inclusion of a real balance effect upon expenditure, showed that even in Sargent's dynamic model, after full adjustment, Mundell's comparative static real balance effect is preserved, whereby a change in anticipated inflation permanently affects the real interest rate.

The proposition is advanced in this paper that the Fisher equation ought to be viewed as a reduced-form relationship derivable from a simple general equilibrium model. This model should allow for the impact of a number of factors which affect the influence of a change in anticipated inflation upon the nominal rate of interest. These factors include taxes on interest earnings; induced changes in income and employment which may accompany a change in anticipated inflation; real balance effects; and the size of the interest elasticity of the demand for money. To account for these factors we introduce a macro-economic model which determines the impact of changes in expected inflation upon the nominal interest rate. The closed economy model will consist of equilibrium conditions in commodity, money, and labor markets with the bond market eliminated by Walras' Law. The money market is expressed in stock equilibrium terms.[3]

We shall entertain the Darby modification of the Fisher hypothesis which states in effect that the after-tax real rate of interest is independent of the level of anticipated inflation. The relationship between

[3]In a recent paper Lewis Johnson derived a result that the real rate is dependent upon the level of anticipated inflation. As in earlier studies, this requires both a real balance effect and a nonzero interest elasticity of money demand. Johnson also introduced a flow equilibrium condition in the money market, "momentary equilibrium," which results in a solution for the relationship between the level of anticipated inflation and the nominal and/or real interest rate dependent upon the speed at which individuals adjust money demand in response to money market (stock) disequilibria. Such a model permits accommodation of the "anticipation effect" whereby individuals anticipating inflation and desiring to cut holdings of real money balances will also realize that their desires will be to some extent satisfied, given no action at all, insofar as anticipated inflation does materialize. Of course it must be assumed that the rate of growth of the nominal money supply is known as well. While Johnson's model is both innovative and promising, it does require explicit knowledge about the speed with which money markets adjust. He in fact concluded that his model assumes "implausible" dynamics whereby portfolio adjustment takes time while goods markets adjust instantaneously. In order to keep a model in a form which involves only parameters for which estimates are readily available, we shall retain the stock equilibrium formulation in the money market. This formulation is equivalent to the flow formulation, as Johnson notes, if it is assumed that instantaneous stock adjustment always keeps flow demand at zero. Later we shall show that the model we propose is stable despite the assumption of instantaneous adjustment in the money market.

ıe Fisher hypothesis and the Darby hy- othesis and their implications can be seen learly from the following. Let i be the ominal interest rate, r the real interest ate, and π the expected rate of inflation. et r^* be the after-tax real interest rate, nd τ be the marginal tax rate on interest ıcome. The Fisher hypothesis begins with ıe relationship

$$i = r + \pi \tag{1}$$

nd states that $di/d\pi = 1.0$, given $dr/d\pi = 0$. he Darby hypothesis begins with the rela- onship

$$r^* = i(1 - \tau) - \pi \tag{2}$$

r, alternatively,

$$i = \frac{r^* + \pi}{(1 - \tau)} \tag{3}$$

ıd states that $di/d\pi = 1/(1 - \tau)$ given $r^*/d\pi = 0$.[4] We shall incorporate equation ı) into the general equilibrium model to be eveloped below.

Equilibrium conditions for our model are ated as follows (with i expressed by equa- on (3)):[5]

$$\begin{aligned} I(r^*) - S(Y(N), M/P) &= 0 \\ L(Y(N), i, P) - M &= 0 \\ PY'(N) - W(P, N^s) &= 0 \end{aligned} \tag{4}$$

The first equation sets real investment I, expressed as a function of the real after-tax rate of return, equal to real savings which is determined by real income and the level of real money balances M/P. Real income, equal to real output, is determined by the quantity of labor employed and is written as $Y(N)$. The second equation sets nominal money demand, $L(Y(N), i, P)$, equal to nominal money supply M. The third equa- tion sets the money wage offered, or the value of the marginal product of labor, $PY'(N)$, equal to the money wage de- manded, expressed as a function of the price level P, and the quantity of labor supplied N^s. Thus N^s is set equal to N, the quantity of labor demanded. Employment and output rise when the money wage of- fered to labor exceeds the current money wage demanded and conversely. The level of anticipated inflation and the money sup- ply are exogenous variables. The first two equations set quantities equal with price adjustments implicitly assumed to clear markets while the labor market equation sets nominal wages demanded equal to nominal wages offered, with adjustments in the quantity of labor implicitly assumed to clear that market. The dynamics of this system will be given explicit consideration in the Appendix.

The system of equations in (4) is differ- entiated and the coefficients expressed in elasticity form. This permits the use of "ball park" parameter values. Some nu- merical analysis is necessary since a priori proximity of $di/d\pi$ to $1/(1 - \tau)$ is an im- portant issue here. Letting Σ_{xy} be defined as "the elasticity of x with respect to y" we summarize signs and ball park figures (in parentheses) for the relevant parameters:[6]

[4]Here we follow Darby and ignore the effect of pos- sible taxes on capital gains which could materialize if commodity prices actually rise at the expected rate. The effect suggested by Darby will follow as long as the marginal tax rate on income lies above that on capital gains. For example, setting $\tau = 0.5$ and the tax rate on capital gains $\tau_K = 0.25$ would imply rewriting (3) as

$$i = \frac{r^*}{1 - \tau} + \frac{1}{1 - \tau_c}\pi \tag{3'}$$

where $\tau_c = (\tau - \tau_K)/(1 - \tau_K) = .33$.

[5]The commodity market is cleared by employing the simplest possible formulations of investment and sav- ings behavior. The term r^* could be included in the savings function with the effect that changes in r^* pro- duce larger changes in excess demand or supply in the commodity markets. The same impact can be intro- duced by varying the sensitivity of investment with re- spect to r^*. The impact of such variation is investigated in the Appendix. Note that in (4), it is assumed that tax proceeds are employed to retire government debt. Explicit government expenditure is omitted from the commodity market equation for simplicity. Its inclu- sion would not affect our major conclusions below.

[6]Our results are unaffected if we allow money de- mand to depend upon after-tax nominal interest, $i^* = i(1 - \tau)$. There of course exist many estimates of the parameter values employed here. Some will prefer to insert values which they consider more likely. This

TABLE 1—COMPARATIVE STATIC PROPERTIES OF A SIMPLE MACRO MODEL

	$\beta = (1 - \Sigma_{WP}) > 0$	$\beta < 0$	$\beta = 0$
$\frac{M}{P}\frac{dP}{dM}$	+	+	+ (1.0)
$\frac{\pi}{P}\frac{dP}{d\pi}$	+	+	+
$\frac{di}{d\pi}$	less than $1/(1-\tau)$	$\lesseqgtr 1/(1-\tau)$ as $\beta \gtreqless \Sigma_{Sm}$	less than $1/(1-\tau)$
$\frac{M}{i}\frac{di}{dM}$	−	+	0
$\frac{M}{N}\frac{dN}{dM}\left(=\frac{M}{Y}\frac{dY}{dM}\right)$	+	−	0
$\frac{\pi}{N}\frac{dN}{d\pi}\left(=\frac{\pi}{Y}\frac{dY}{d\pi}\right)$	+	−	0

investment: $\Sigma_{Ir^*} < 0 \quad (-0.4)$

savings: $\Sigma_{SY} > 0 \quad (1.0)$

$\Sigma_{Sm} < 0 \quad (-0.2)$

$m = M/P$

money demand: $\Sigma_{LY} > 0 \quad (1.0)$

$\Sigma_{Li} < 0 \quad (-0.5)$

$\Sigma_{LP} > 0 \quad (1.0)$

output:[7] $\Sigma_{YN} > 0 \quad (1.0)$

$\Sigma_{Y'(N)N} = 0$

wage demand: $(0 \leq \Sigma_{WP} \leq 1.69)$

$\Sigma_{WN} > 0 \quad (1.0)$

The parameter of most interest here is the elasticity of money wages demanded with respect to price Σ_{WP}, upon which we have placed a range estimate running from 0 to 1.69. Neither number is arbitrary. The lower limit reflects downward rigidity of money wages. The upper limit represents the maximum value of Σ_{WP}, given other parameter values already specified, that is consistent with dynamic stability (discussed in the Appendix) of the system described by (4). The term Σ_{WP} is not expected to be constant. During the initial stages of a rise in the rate of inflation, institutional rigidities associated with wage contracts or slowness in identifying the true process describing the behavior of prices lead us to expect $\Sigma_{WP} < 1$.[8] For the same reasons, during the initial stages of a fall in the rate of inflation or during the period when (previously reduced) real wages are being restored, we expect $\Sigma_{WP} > 1$. During steady inflation as the true process describing the behavior of prices comes to be correctly perceived we expect Σ_{WP} to approach unity. As a matter of fact, estimates of Σ_{WP} reported in the survey of the Phillips curve literature by Rutledge vary widely over time and location, with values reported ranging from about 0.2 to unity. We shall investigate the significance of changes in Σ_{WP} for our results once the comparative static properties of (4) have been determined.

can easily be done, but it will be discovered that the nature of our findings is not largely altered by changes to other parameter values which lie within the range frequently estimated. (See the Appendix.)

[7] We assume constant returns to scale.

[8] These factors are emphasized by William Poole in a discussion of phenomena which may lead to observation of events such as business cycles or real effects of nominal disturbances which are inconsistent with fully rational expectations. In addition, as noted by Milton Friedman, higher rates of inflation tend also to be more volatile. Therefore prices in the future become more difficult to predict from prices now and *ex post* values of Σ_{WP} will be more likely to differ from unity when rates of inflation are high and volatile.

Differentiating (4) and converting into elasticity form gives results summarized in Table 1. The sign of the relationship between percentage changes in exogenous M and π, and the percentage change in each of the endogenous variables P, i, and N is given to acquaint the reader with the general properties of the model under alternative values of β, defined equal to $1 - \Sigma_{WP}$. The impact of anticipated inflation on interest is given simply as $di/d\pi$ for comparison with the value obtained for this derivative under the Darby hypothesis. An explicit expression for $di/d\pi$ implied by our model is given below in (5). Recall that given percent changes of M and π will produce percent changes of N identical to those of Y since $\Sigma_{YN} = 1.0$.

Our primary interest is in $di/d\pi$. The other results will be given some further consideration below. First, it can readily be seen from Table 1 that values of $di/d\pi$ smaller than that anticipated by the Darby hypothesis will arise for nearly all possible values of $\beta = (1 - \Sigma_{WP})$. Further, an elasticity of money wages demanded with respect to inflation of greater than unity implies a reduction in real output and employment when the money stock is increased.[9]

Taking Σ_{YN}, Σ_{SY}, and Σ_{WN}, all equal to unity, we may write:[10]

$$(5)\quad \frac{di}{d\pi} = \frac{1}{(1-\tau) + \dfrac{\Sigma_{Li}(\beta - \Sigma_{Sm})}{(i/r^*)\Sigma_{Ir^*}(\beta + 1)}}$$

Equation (5) gives the Mundell-Sargent-Visco result since, given $\tau = 0$, $di/d\pi = 1.0$ requires setting Σ_{Li} equal to zero.[11] Thus $\beta = 0$ is equivalent to a full employment assumption in a model where real wages are constant and equal to the marginal product of a given (full employment) quantity of labor. Mundell did fail to note, however, that his result whereby $di/d\pi < 0$ given the real balance effect also required $\Sigma_{Li} < 0$, though this condition was implicit, as we have noted, in the negatively sloped LM curve appearing in his Metzler-type diagram. His result gives $di/d\tau = 0$ if $\Sigma_{Ir^*} = 0$.

Crude as they may be, the ball park parameter values permit us to suggest some limits for values of $di/d\pi$ under different assumptions about i/r^* and β as well as τ. One thing which characterizes empirical estimates of $di/d\pi$ is variety, and some investigators like Gibson and Lahiri report evidence of structural changes or breaks in estimated values of $di/d\pi$, particularly around 1960. First we note that, given Σ_{WN}, Σ_{YN}, $\Sigma_{SY} = 1.0$, if $\beta > \Sigma_{Sm} > -1$:

$$(6)\quad \frac{d}{d\beta}[di/d\pi] = \frac{(-i/r^*)\Sigma_{Ir^*}\Sigma_{Li}[1 + \Sigma_{Sm}]}{|A|^2} < 0$$

$$(7)\quad \frac{d}{d(i/r^*)}(di/d\pi) = \frac{\Sigma_{Li}\Sigma_{Ir^*}(\beta - \Sigma_{Sm})(\beta + 1)}{|A|^2} > 0$$

where $|A| \equiv [(i/r^*)\Sigma_{Ir^*}(1 - \tau)(\beta + 1) + \Sigma_{Li}(\beta - \Sigma_{Sm})] < 0$

[9]Some other results given in Table 1 are worth noting. When $\beta = 0$ the natural rate hypothesis is satisfied by the classical result. the rate of increase in the price level is identical to the rate of increase in the money supply (the "1" under $\beta = 0$), while output and employment are independent of monetary growth. The Fisher/Darby hypothesis is contradicted even in the classical case where $\beta = 0$ due to the real balance effect, as was noted by Mundell. The result whereby $(\pi/N)(dN/d\pi) > 0$ is exactly that hypothesized and measured by Fisher (1926) in his "discovery" of the Phillips curve whereby employment too is termed a "dance of the dollar." Of course Fisher is appropriately skeptical about the ultimate impact of rising inflation upon employment, noting that, after a rise in inflation, "Employment is then stimulated—for a time at least" (p. 498).

[10]The labor market-clearing equation in (4) has money wages demanded dependent upon the price level and the quantity of labor supplied. Thus our assumption $\Sigma_{WN} = 1.0$ is equivalent to an assumption of unitary elasticity of labor supply with respect to money wages at a given price level or $1/\Sigma_{WN} = \Sigma_{NW} = 1.0$. Such an assumption, which amounts to unitary elasticity of labor supply with respect to real wages, seems reasonable enough and serves to reduce the complexity of comparative static solutions to the model.

[11]The only other possibility would have a negative β equal in absolute value to Σ_{Sm} given Σ_{WN}, Σ_{SY}, and $\Sigma_{YN} = 1.0$.

Now, as expected inflation rises we expect Σ_{WP} to rise thereby lowering β and, given (6), in turn raising $di/d\pi$. At the same time, a rise in π raises $i/r^* = (1/(1 - \tau))(1 + \pi/r^*)$ as long as π increases faster than r^*. This in turn adds to the positive impact upon $di/d\pi$ arising from an increase in Σ_{WP} since (7) carries a positive sign.

To place a lower limit on $di/d\pi$, set $\Sigma_{WP} = 0$ ($\beta = 1.0$). If initially $\pi = 0$, $i/r^* = 1/(1 - \tau)$. We set $\tau = 0.33$ consistent with the marginal tax rate on nominal interest suggested by Darby to be consistent with observed differentials between taxable and tax-free returns on bonds. These parameter values along with those already given imply $di/d\pi = 0.857$.

Alternatively, suppose $r^* = \pi = 4.0$ percent and $\Sigma_{WP} = 1.0$ ($\beta = 0$). Given these assumptions, $di/d\pi = 1.333$. It is also clear from (5) that as β approaches Σ_{Sm}, in this case -0.2, $di/d\pi$ approaches $1/(1 - \tau)$. Finally, it is easy to account for estimates of $di/d\pi$ close to unity by selecting values between the extremes already mentioned. Given $i/r^* = 2.0$ and $\tau = 0.33$, a value of $\beta = 0.695$ ($\Sigma_{WP} = 0.305$) gives a value of $di/d\pi = 1.0$ since these values satisfy

$$\tau = \frac{\Sigma_{Li}(\beta - \Sigma_{Sm})}{i/r^* \, \Sigma_{Ir^*}(\beta + 1)} \tag{8}$$

which, given equation (5), is a necessary and sufficient condition for $di/d\pi = 1.0$.

As simple as our model is and as crude as our parameter estimates may be, it is by now evident that it is an easy matter to employ both to account, first, for most of the estimated values of $di/d\pi$ and, second, for changes in the value of $di/d\pi$ over time, given changes over time in the value of Σ_{WP}. Finally we can show that the existence of estimates of $di/d\pi$ close to unity do not confirm the Fisher hypothesis in any general sense and in particular do not imply after-tax real returns that are independent of the level of anticipated inflation.

Returning to Table 1, it is worthwhile to consider the general economic situations which could be expected to accompany changing values of $di/d\pi$. Lower values of $di/d\pi$ would be expected to arise when Σ_{WP} is relatively low (less than one so that $\beta > 0$) due say to a recent rise in the rate of inflation. Given such circumstances an exogenous increase in π will raise prices, real output, employment, and nominal interest rates. If π continues to rise, given that wages begin to respond more promptly to increases in actual inflation, β will fall while i/r^* rises and if β becomes negative, there may arise a period of rising inflation and falling real output and employment, accompanied by rising nominal interest rates. In fact β describes the ratio of the impact upon real output (employment) and prices arising from monetary expansion or a rise in anticipated inflation based on the model given by (4). The implied mix in the relative response of real output (employment) and inflation to monetary growth or a rise in anticipated inflation goes from a maximum value of 1.0 ($\beta = 1.0$, $\Sigma_{WP} = 0$) downward to a minimum value of -0.69 ($\Sigma_{WP} = 1.69$).[12] The case where $\beta = 0$ ($\Sigma_{WP} = 1.0$) is consistent with the natural rate hypothesis (a vertical Phillips curve) which in turn reflects the absence of money illusion implicit in a value of $\Sigma_{WP} = 1.0$. In addition where β is falling for reasons cited above, $di/d\pi$ can be expected to rise, suggesting that real effects follow from changes in nominal M or π, when the inflation rate forecast is wrong. Such real effects vary with the degree of inaccuracy of inflation forecasts. Under the assumption however that inflation forecasts are correct in the long run, the steady-state value of Σ_{WP} is unity and $\beta = 0$. Our results are then consistent with the notion that an unemployment inflation tradeoff is a short-run phenomenon.

The economics behind the results just discussed are straightforward. The rise in expected inflation produces an excess demand for commodities as investment rises due to an initially depressed real rate of interest. Simultaneously there occurs an excess supply of money in the face of antici-

[12]Based on the requirement that β lie above -0.69 for the model in (4) to be stable. This result is derived in the Appendix.

pated depreciation of money in terms of goods, caused by the rise in π and the initial rise in the nominal interest rate. In short, anticipated inflation leads to a desire to convert financial to real assets. In a full employment model with no real balance effects the new equilibrium in the commodity sector requires a restoration of the original after-tax real rate which in turn requires $di/d\pi = 1/(1-\tau)$. With real balance effects, the higher price level resulting from excess commodity demand lowers real balances, thereby shifting out the savings schedule and requiring a lower after-tax real interest rate to reequilibrate the commodity sector. Of course, if the interest elasticity of money demand is zero, no excess supply condition ever arises in the monetary sector. Even with a real balance effect there is no impact upon the after-tax real interest rate since no reduction in money demand is available to increase commodity demand. The price level remains constant and i simply rises by $\pi/(1-\tau)$.

Once the assumption of fixed output is dropped and attention is paid to the impact of price level changes upon real wages, employment, and output, the potential for an impact of changes in expected inflation upon the real rate of interest is further enhanced. If $\Sigma_{WP} < 1$ $(\beta > 0)$, then the changes described above are accompanied by a rise in real income (output). As a result savings are increased and the new equilibrium real interest rate can be still lower in response to a rise in π. The condition $\Sigma_{WP} = 1.0$ freezes real income which simply returns us to the full employment case of Mundell, while $\Sigma_{WP} > 1$ means that real income falls given a rise in π. This in turn means that $dr^*/d\pi$ will be $\lesseqgtr 0$ $(di/d\pi \lesseqgtr 1/(1-\tau))$ depending on whether the fall in real balances results in an increase in savings that is greater than, equal to, or less than the decrease in savings caused by a fall in real income.

Finally, it is interesting to note that the results presented in Table 1 rationalize the Gibson Paradox, whereby a higher price *level* is associated with higher nominal interest rates, in cases were Σ_{WP} exceeds unity. In such cases monetary expansion will result in inflation and higher nominal interest which will in turn cause high levels of prices and interest rates to appear simultaneously.

II. Measurement of the Impact of Anticipated Inflation on Nominal Interest

Ever since Fisher uneasily concluded in 1930 that "when prices are rising, the (nominal) rate of interest tends to be high *but not so high as it should be to compensate for the rise*" (p. 43, emphasis added), empirical investigations into the effect upon nominal interest rates of changes in anticipated inflation rates have generally ended with some *ad hoc* theorizing, either on why it is that nominal interest rates do not rise by the full amount of an increase in anticipated inflation, or why it is that the measured impact of changes in inflationary expectations upon nominal interest rates is not constant over time.[13]

Frequently the problem is found to be one of properly modeling and measuring just how it is that people form their expectations about future rates of inflation. The implication of the Fisher legacy is a search for measures of anticipated inflation that can predict nominal interest rates with a regression coefficient of unity. As a consequence we find studies like that of Lahiri employing "weighted," "adaptive," "extrapolative," and "regressive" expectations to model the formulation of expected inflation rates along with one-, two- or three-stage least squares estimates of the impact of such weights upon nominal interest rates. Even with all of these variants the results do not support the hypothesized impact upon nominal interest rates arising from changes in expected rates of inflation, however measured, and there appears to be a structural break in the data around 1960. It is our contention here that neither of these results is at all surprising and that the prior beliefs

[13]See articles surveyed by Richard Roll, and more recently, articles by Gibson; Lahiri; Carr, Pesando, and Smith.

generated by the Fisher hypothesis can and have in some cases seriously misled empirical investigators.

Despite the rather extensive efforts of theorists investigating the Fisher hypothesis, empirical researchers have in some cases found theorists easy to ignore, particularly on the question of constancy of the real rate. Gibson, citing Reuben Kessel and Armen Alchian and Mundell, simply assumes $dr/d\pi = 0$ because "There is as yet no theoretical consensus on the relationship between the real rate and the expected rate of inflation" (p. 855). This might be ignored if he did not conclude later on that his results "...lend support to the hypothesis that the real rate of interest is not affected by price expectations over a six month period..." (p. 863).[14]

Lahiri appears to recall late in his article the significance of his initial assumption that the real rate is independent of anticipated inflation when he writes somewhat cryptically: "Though some effects of price expectations on [the] real rate can never be overemphasized, I did not pursue that point in this paper" (p. 130).

The Fisher legacy has been carried a step further by Fama. Based on the joint hypothesis that expectations are rational and that the real rate of interest is constant, Fama concludes that one cannot reject the hypothesis that "all variation through time in one- to six-month nominal rates of interest mirrors variation in correctly assessed one- to six-month expected rates of change in purchasing power" (p. 282). Fama in effect requires that the Fisher *hypothesis*, which predicts a unit impact upon nominal interest rates of changes in expected inflation, be empirically verified if we are to conclude first, that expectations are rational and second, that the real rate is constant in general and, in particular, is independent of changes in the expected rate of inflation.[15]

Fama's innovative empirical approach which requires that both market efficiency (in the sense of rational expectations) and a constant real rate be included in the null hypothesis, carries with it the danger that the hypothesis of market efficiency may erroneously be rejected if one clings to the validity of the hypothesis that real rates are constant. In a paper modifying Fama's conclusions, Charles Nelson and G. William Schwert (N-S) argue that Fama's test could not have been expected to produce rejection of his joint hypothesis. More powerful tests performed by N-S do result in rejection of that hypothesis. They choose to conclude, based upon copious evidence of market efficiency, much of it produced by Fama himself, that their results suggest rather that the real interest rate is not constant. We have argued that if the after-tax real rate of interest depends upon the level of anticipated inflation, efficient markets imply that the Fisher hypothesis *cannot* be fulfilled but that one may, given the validity of the Darby hypothesis, erroneously infer that the Fisher hypothesis holds.

The prior belief that $di/d\pi = 1.0$ implies independence of the real rate from anticipated inflation really involves a compound error, given a general equilibrium model coupled with the Darby hypothesis of independence of the real after-tax rate from anticipated inflation. To see this consider first the implication of the general equilibrium model for the value of $di/d\pi$ given only the Fisher hypothesis and investment determined by the real rate unadjusted for taxes, r. In this case

[14]To be fair it should be noted that Gibson is here commenting on a finding that for some of his results $di/d\pi$ lies close to unity which, ignoring tax effects, *could* imply $dr/d\pi$ close to zero. But many of his results do not show $di/d\pi$ close to unity and as Gibson himself notes $di/d\pi = 1$, while consistent with $dr/d\pi = 0$, is also consistent with a world in which "positive (negative) effects on the real rate are exactly matched by underadjustment (overadjustment) of nominal rates" (p. 855).

[15]It is important to distinguish clearly between the Fisher *hypothesis* stated here and the Fisher *finding* that indeed the nominal interest rate does *not* appear to change by the full amount of a change in the expected rate of inflation which is measured in turn by a distributed lag on past and current observed rates of inflation.

$$(5') \qquad \frac{di}{d\pi} = \frac{1}{1 + \dfrac{\Sigma_{Li}(\beta - \Sigma_{Sm})}{(i/r)\Sigma_{Ir}(\beta + 1)}}$$

which becomes unity under conditions identical to those which make (5) equal to $1/(1 - \tau)$. Given (5′) we are not surprised to find estimates of the impact of changes in π upon i lying below unity. However, note that if $\beta = \Sigma_{Sm}$, and negative effects of falling income on saving arising from a rise in real wages (given $\Sigma_{WP} > 1.0$ and $\beta < 0$) cancel positive effects of lower real balances on saving, the real rate may be unaffected. But this is a rather special case, likely to appear only in a period of rapidly decelerating inflation and/or past reductions in real wages. In any case, should the estimated impact of a rise in anticipated inflation upon the nominal interest rate be unity, we could, given (5′), correctly infer under the Fisher hypothesis, a constant real rate.

The real difficulty associated with the prior belief that $di/d\pi = 1.0$ implies independence of the real rate from anticipated inflation arises under the Darby hypothesis, given that we take investment to depend upon the after-tax real rate. If one looks back at our equation (5), it is seen that its right-hand side exceeds the r.h.s. of (5′) when $\tau > 0$ and $\Sigma_{Ir^*} = \Sigma_{Ir}$.[16] As we have already noted, the r.h.s. of (5) can assume a value of unity, given our parameter values when $\Sigma_{WP} = .30$. Should this result coincide with the prior beliefs associated with the Fisher hypothesis, one would wrongly conclude, given the Darby hypothesis and $I = I(r^*)$, that the real rate is unaffected by anticipated inflation when in fact such a result implies that neither the after-tax real rate nor the real rate displays such independence.[17] Thus Fama's estimate of 0.98 for the impact of anticipated inflation on nominal interest supports neither independence of the real rate nor of the after-tax real rate from anticipated inflation.

Tests of the Darby hypothesis conducted on Canadian data by Carr, Pesando, and Smith which were termed "inconclusive" by their authors in fact produced results quite consistent with those predicted by our equation (5). A wide variety of formulations to represent expected inflation produced statistically significant estimates of $di/d\pi$ running from 0.86 to 1.34 with a mean value of 1.03 and a standard deviation of 0.12 employing interest rates on instruments with a range of maturities from ninety days to ten years.[18]

Carr, Pesando, and Smith given a prior belief based on the Darby hypothesis of an after-tax *real* rate independent of anticipated inflation that $di/d\pi = 1/(1 - \tau)$ were led to the "inconclusive" view of their results. We would argue simply that they entertained an incorrect prior belief.

III. Concluding Remarks

It is our view that empirical investigators of the effects of anticipated inflation have not been well served by prior beliefs based either on the Fisher hypothesis or the Darby hypothesis. The Fisher hypothesis has tended to serve as a criterion for the validity of measures of anticipated inflation for those investigators who search for the measure which results in an estimate of $di/d\pi$ close to unity. Some investigators like Fama have estimated values of $di/d\pi$ that are in fact close to unity and, we contend, have wrongly inferred independence of the real rate from anticipated inflation. In-

[16]Note that i/r^* will exceed i/r since $r - r^* = i\tau$.

[17]This follows since independence of the after-tax rate requires $di/d\pi = 1/(1 - \tau)$ (the Darby hypothesis). Also, when the r.h.s. of (5) equals unity, and we ignore the $i\tau$ term for $r - r^*$, it implies the r.h.s. of (5′) equal to $1/(1 + \tau)$ which contradicts the Fisher hypothesis.

[18]Based on data running from 1959–71, there were thirty such estimates deemed acceptable by Carr, Pesando, and Smith with ten others rejected due to some wrong or insignificant signs on estimated coefficients. Since the authors admit to being unable to reconcile all the different estimates of $di/d\pi$, the rather crude expedient of characterizing their results by simply calculating a mean and standard deviation for all estimates obtained does not seem an inappropriate way to characterize the state of such estimates. Implications for results of Carr, Pesando, and Smith arising from the openness of the Canadian economy from which data were drawn are considered in Makin.

vestigators like Carr, Pesando, and Smith, who have employed the Darby hypothesis as a basis for their prior beliefs about $di/d\pi$, have been misled to view their results as inconclusive when in fact they are quite consistent with results predicted by a general equilibrium approach to the effects of changes in anticipated inflation.

More generally, it is our view that, like the Philips curve, neither the Fisher hypothesis nor the Darby hypothesis represents an isolated phenomenon, but rather should be viewed as a reduced-form relationship derivable from a set of structural equations which compose a reasonably comprehensive macro-economic model. Viewed in this way the results obtained by empirical investigations of both hypotheses are at once consistent with expectations based upon theory and inconsistent with the notion that the real rate of interest, before or after taxes, is independent of the level of anticipated inflation. Insofar as instruments of monetary and fiscal policy have, under rational expectations, a direct effect on the rate of anticipated inflation, then the conclusion implied by our results also implies that real variables are directly affected by such instruments.

Finally, our results suggest that the role played by income effects, as opposed to that played by real balance effects, in affecting the impact of anticipated inflation upon nominal interest may vary over time due to changes in the elasticity of money wages demanded with respect to prices. A shift in that parameter may help to explain the discovery by Lahiri, Gibson, and William Yohe and Denis Karnosky, of a break occurring about 1960 in the measured impact of anticipated inflation on nominal interest. More generally, these and other empirical investigators of the Fisher hypothesis, and recently, of the Darby hypothesis, may be obtaining a wide range of estimates of the impact of changes in anticipated inflation upon nominal interest, not because of any wide variation in the adequacy of different proxies employed to measure anticipated inflation, but rather because of the simple fact that they are attempting to measure a coefficient which varies randomly over time.

Appendix A: Constraining Σ_{WP} with the Correspondence Principle

Here we explore the dynamic properties of the model given in (4) as a means to place some lower limit on values of β by employing Samuelson's Correspondence Principle. It will also be of interest to consider the behavior of a system in which two markets (commodities and money) are assumed to adjust via price adjustments while a third (labor) is assumed to adjust via quantity adjustments. We have:

$$\text{(A1)}\quad \begin{aligned} dP/dt &= K_1[I(r^*) - S(Y(N), M/P)] \\ di/dt &= K_2[L(Y(N), i, P) - M] \\ dN/dt &= K_3[PY'(N) - W(P,N)] \end{aligned}$$

The characteristic equation which determines the solution of the above system of linear differential equations in the neighborhood of equilibrium is (letting $K_i = 1.0$ $i = 1, 2, 3$):

(A2)

$$\begin{vmatrix} \Sigma_{Sm} - \lambda & (1-\tau)(i/r^*)\Sigma_{Ir^*} & -1.0 \\ 1.0 & \Sigma_{Li} - \lambda & 1.0 \\ \beta & 0 & -1.0 - \lambda \end{vmatrix} = 0$$

Local stability requires that the roots of (A2) be negative. Equation (A2) may be written out as a cubic equation:

$$\text{(A3)}\quad a_0\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0$$

where

$$a_0 = 1$$

$$a_1 = (-\Sigma_{Sm} - \Sigma_{Li} + 1) > 0$$

$$a_2 = -\Sigma_{Li} + \Sigma_{Sm}(\Sigma_{Li} - 1) - (1-\tau)(i/r^*)\Sigma_{Ir^*} + \beta \gtrless 0 \quad \text{as } \beta \gtrless -1.6$$

$$a_3 = -|A| \gtrless 0 \text{ as } \beta \gtrless -0.69$$

The roots of equation (A3) will be negative

TABLE 2—SENSITIVITY OF $di/d\pi$ TO PARAMETER CHANGES

Parameter Range	Range of $di/d\pi$					
	$\beta = 1.0$		$i/r^* = 1.5$	$\beta = 0$		$i/r^* = 3.0$
$\Sigma_{Sm} = -0.1$ to -0.3	0.889	to	0.827	1.411	to	1.262
$\Sigma_{Li} = -0.2$ to -0.8	1.153	to	0.682	1.428	to	1.249
$\Sigma_{Ir^*} = -0.1$ to -0.7	0.375	to	1.050	1.00	to	1.400

if $a_1 > 0$; $a_2 > 0$; $a_3 > 0$; $a_1a_2 - a_0a_3 > 0$. A sufficient condition to satisfy all four conditions given the ball park parameters employed above (given $\tau = 0.33$ and $i/r^* = 3.0$) is $\beta > -0.69$ or $\Sigma_{WP} < 1.69$.

If the money market is assumed to adjust instantaneously so that stock equilibrium and flow equilibrium conditions are identical, the characteristic equation for the system becomes a quadratic equation which has the same sufficient condition for negative roots (stability) as the cubic equation.

APPENDIX B: IMPACT OF CHANGES IN PARAMETER VALUES ON $di/d\pi$

As already noted in footnote 6, the results obtained for $di/d\pi$, while not radically altered, will be affected by changes in parameter values. Assuming parameters employed in the text except as noted otherwise, range estimates are given for $di/d\pi$ in Table 2. The results suggest first, that lower estimates of $di/d\pi$ obtained by Gibson are consistent with either a higher interest elasticity of money demand with respect to nominal interest or a lower elasticity of investment with respect to after-tax real interest. Second, adjustments of Σ_{Ir^*} and Σ_{Li} in the same direction would tend to offset each other. Third, the general impression obtained from Table 2 is one of relative stability for estimated values of $di/d\pi$ with sensitivity to changes in Σ_{Ir^*} being the greatest. Finally, sensitivity of $di/d\pi$ to changes in Σ_{Li} and Σ_{Ir^*} falls as the rate of anticipated inflation and Σ_{WP} rise (β falls). In view of the persistent actual and therefore, anticipated inflation in most countries over recent years, the ranges associated with $\beta = 0$ are probably more typical and of course will be likely to hold in the long run given constancy of real wages.

APPENDIX C: IMPACT OF CAPITAL GAINS TAXES ON $di/d\pi$

If $\tau \neq \tau_K$ we obtain the result:

$$(5'') \quad \frac{di}{d\pi} = \frac{1}{\dfrac{1-\tau}{1-\tau_K} + \dfrac{\Sigma_{Li}(\beta - \Sigma_{Sm})}{(1-\tau_K)(i/r^*)\Sigma_{Ir^*}(\beta + 1)}}$$

Given parameters associated with $\beta = 1.0$ (see text) and values of $\tau = 0.5$ and $\tau_K = 0.25$, $di/d\pi = 0.750$. If $\beta = 0$, the comparable value for $di/d\pi$ becomes 1.285. As is obvious from (5″) implied values of $di/d\pi$ fall as the difference between τ and τ_K rises.

REFERENCES

J. Carr, J. E. Pesando, and L. B. Smith, "Tax Effects, Price Expectations and the Nominal Rate of Interest," *Econ. Inquiry*, June 1976, *14*, 259–69.

M. R. Darby, "The Financial and Tax Effects of Monetary Policy on Interest Rates," *Econ. Inquiry*, June 1975, *13*, 266–76.

E. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

Irving Fisher, "A Statistical Relation Between Unemployment and Price Changes," *Int. Labour Rev.*, June 1926, *13*, 785–92; reprinted as "I Discovered the Phillips Curve," *J. Polit. Econ.*, Mar./Apr. 1973,

81, 496-502.

———, *The Theory of Interest*, New York 1930.

M. Friedman, "Nobel Lecture: Inflation and Unemployment," *J. Polit. Econ.*, June 1977, *85*, 451-72.

W. E. Gibson, "Interest Rates and Inflationary Expectations," *Amer. Econ. Rev.*, Dec. 1972, *62*, 854-65.

L. Johnson, "Inflationary Expectations and Momentary Equilibrium," *Amer. Econ. Rev.*, June 1976, *66*, 395-400.

R. A. Kessel and A. A. Alchian, "Effects of Inflation," *J. Polit. Econ.*, Dec. 1962, *70*, 521-37.

K. Lahiri, "Inflationary Expectations: Their Formation and Interest Rate Effects," *Amer. Econ. Rev.*, Mar. 1976, *66*, 124-31.

J. H. Makin, "Anticipated Inflation and Interest Rates in an Open Economy," paper no. 77-1, Instit. Econ. Res., Univ. Washington, Jan. 1977.

R. A. Mundell, "Inflation and Real Interest," *J. Polit. Econ.*, June 1963, *71*, 280-83.

C. R. Nelson and G. W. Schwert, "On Testing the Hypothesis that the Real Rate of Interest is Constant," *Amer. Econ. Rev.*, June 1977, *67*, 478-86.

W. Poole, "Rational Expectations in the Macro Model," *Brookings Papers*, Washington 1976, *2*, 463-505.

R. Roll, "Interest Rates on Monetary Assets and Commodity Price Index Changes," *J. Finance*, May 1972, *27*, 251-78.

J. Rutledge, "The Unemployment Inflation Tradeoff: A Review Article," *Claremont Econ. Papers*, No. 141, July 1975.

T. J. Sargent, "Anticipated Inflation and Nominal Interest," *Quart. J. Econ.*, May 1972, *86*, 212-25.

I. Visco, "Inflation and the Rate of Interest," *Quart. J. Econ.*, May 1975, *89*, 303-10.

W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes," *Fed. Reserve Bank St. Louis Rev.*, Dec. 1969, *51*, 19-36.

A Calculus Approach to the Theory of the Core of an Exchange Economy

By LEIF JOHANSEN*

The theory of the shrinking of the core of an exchange economy to the competitive equilibrium (or set of equilibria) when the number of participants increases is one of the most important and interesting contributions to general equilibrium theory in recent decades, and ought to become part of standard courses in economic theory. It is important to have an exposition of this idea which appears as a simple and natural extension of the tools of analysis familiar to most students of economics. The purpose of the present paper is to make an attempt at such an exposition along traditional calculus lines. The paper does not contain results which are new to specialists in the field. In the literature there are, of course, some expositions which point in the direction taken here, but I have not seen the approach spelled out in the way it is done in the sequel. (Some relevant references are given at the end of the paper.)

I. Background and Perspectives

Let me first state very briefly why I consider the result mentioned to be interesting and important. It is then necessary to emphasize the difference between the meaning of the concepts of competitive equilibrium and core allocations.

A competitive equilibrium presupposes the existence of a price system. Under this system individual agents act in isolation in the sense that each of them decides how much to supply and how much to demand of the various commodities on the basis of his own preferences, without making conscious and explicit arrangements with other agents. Each agent considers prices as given in an impersonal way, not subject to bargaining or manipulation through his own supply and demand. We have equilibrium if prices are such that supply and demand for all agents taken together are equal for each commodity. Provided that we have somehow established equilibrium prices in this sense, they solve a complicated multiagent problem by transforming it into a set of rather simple individual decision problems. (It is not necessary for our purpose to go into the problem of how the prices are established and the associated dynamic stability problems.)

A core allocation is defined by an entirely different approach. In this case we consider only a set of agents with initial holdings of commodities who may improve their positions by reallocation, but we do not presuppose the existence of a price system. We start at a more basic level, assuming only that there are possibilities for the agents to communicate and make agreements to reallocate commodities between them—by unilateral gifts, by bilateral exchange, or by some more complicated multilateral exchange arrangement. The individuals are free to form "coalitions" for the purpose of improving the situation for members of the coalition. In our context a coalition is simply a group of agents who agree on a certain reallocation of the initial quantities of goods held by its members. It should be observed that the initial quantities are individually owned, and ownership respected in the sense that nothing can be taken away from an agent without his consent, as part of a voluntary exchange or reallocation. We may now ask whether it is possible to predict the outcome of the exchange or reallocation process in such a system. The "core" gives an answer to this question. It is based on the following observation: Consider an outcome which is feasible in the

*Professor of economics, University of Oslo. I am grateful to a referee for useful remarks and suggestions.

sense that the total amounts of commodities held after the exchange or reallocation are equal to the total initial amounts. This outcome implies a specific bundle of commodities for each agent. If there is at least one group of agents such that these agents could improve their situations by redistributing their own initial holdings instead of agreeing to the proposed outcome, then this outcome will not be realized. It will be "blocked" by the group or coalition mentioned, that is, they will refuse to accept it, because there is another arrangement which they can realize without requiring the cooperation of other agents, and which is better for each of them. It is then natural to ask: Is there a feasible outcome, or a set of outcomes, which will not be blocked by any coalition that can be formed by some agents of the economy (including degenerate coalitions consisting of single agents, and the grand coalition comprising all agents). If such an outcome, or set of outcomes exists, then this is the core.

An outcome belonging to the core is stable in a very important sense, different from the usual dynamic stability concept. It is stable against attempts by individuals and coalitions to find something better, because no possible coalition can do better by refusing to accept the outcome, and instead manage on the basis of the initial holdings of the members of the coalition.

We can now compare the allocation defined by the competitive equilibrium with core allocations. It is a simple matter to show, for exchange economies which we shall consider here, that the competitive equilibrium allocation belongs to the core. The result referred to above as the theory of the "shrinking of the core of an exchange economy to the competitive equilibrium (or set of equilibria) when the number of participants increases" is more striking and also more complicated to prove, and it is to this theme the present paper will be devoted. This connection between competitive equilibria and the core may, in my opinion, give rise to rather far-reaching speculations about economic systems and institutions.

The establishing of a core allocation by means of tentative formations of coalitions of all sizes and compositions, and comparisons of feasible outcomes for the various coalitions, will for an economy with more than a handful of agents represent a large effort in terms of communication and negotiations. In comparison the mechanism of competitive equilibrium is strikingly simple, requiring only individual decisions (when correct prices are given). If an economy has, so to speak, invented the price mechanism, and competitive equilibrium prices have been established, then an enormous organizational problem is solved in an easy manner, and the solution is stable in the sense described above, that is, any group which might contemplate breaking out of the market system will in the end find that it cannot improve its situation by doing so. Furthermore, if the economy consists of a number of agents "approaching infinity," then outcomes corresponding to the set of competitive equilibria are the *only* outcomes which satisfy this kind of stability requirement. I think these considerations go a long way towards explaining why competitive market mechanisms have appeared in almost all corners of the world and, under almost all conceivable circumstances, why they have proved to be so robust, why other arrangements tend to be less permanent, and why attempts to abolish the market mechanism have often failed in the sense that markets reappear unofficially parallel with the official nonmarket system. (These are, of course, sweeping statements which should not be taken too literally. They are meant only as suggestions of the perspectives opened up by a seemingly rather formal and esoteric theory.)

II. Strategy of Reasoning

As already suggested, this paper will treat only exchange economies, although extensions to production economies are possible. The main idea is to get as far as we can by means of simple calculus tools of analysis. We must then assume more of "smoothness" than necessary in more advanced expositions and proofs. In fact, we shall

assume strictly convex preferences, representable by differentiable utility functions. The advantage gained by this is that we can exploit the possibility of approximating a utility function by its tangent in certain neighborhoods.

The strategy of the reasoning is first to limit considerations to Pareto optimal points since both competitive equilibria and core allocations belong to the set of Pareto optima. (The last part of this statement is true because the coalition of all agents would block, in the sense indicated above, any allocation which is not Pareto optimal. This, by the way, points to a limitation of the core theory in the form considered here. Whenever we consider, as we often do in welfare theory, situations which are not Pareto optimal, then we must implicitly assume some sorts of difficulties which prevent the formation of coalitions. It is, however, beyond the scope of this paper to pursue this idea.) Then we consider the various Pareto optimal allocations to see whether there are coalitions which would block them, and we shall find that, for any such allocation which does not belong to the set of competitive equilibria, we can construct such a coalition, that is, prove that the allocation does not belong to the core, provided that the number of agents is sufficiently large. (A certain regularity may be required concerning the way in which the number of agents is made large.)

III. Description of the Economy and Notation

I now introduce the notation necessary to describe the exchange economy to be considered. Let there be M perfectly divisible commodities indexed $i = 1,\ldots, M$ and G "types" of individuals, indexed $j = 1,\ldots, G$. All individuals of the same type have the same initial quantities of the various commodities and the same utility functions. The following notation is also introduced:

N_j = the number of individuals of type j ($j = 1,\ldots, G$).

$\bar{x}_{ij}$ = *initial quantity* of commodity i held by a person of type j ($i = 1,\ldots, M$; $j = 1,\ldots, G$).

x_{ij} = quantity of commodity i held by a person of type j after the exchange ($i = 1,\ldots, M$; $j = 1,\ldots, G$). I call these *final quantities*.

$U_j = U_j(x_{1j},\ldots, x_{Mj})$ = utility function of an individual of type j ($j = 1,\ldots, G$). Assumptions about the utility functions have already been mentioned in Section II above.

$u_{ij} = \partial U_j/\partial x_{ij}$ = marginal utility of commodity i for an individual of type j ($i = 1,\ldots, M$; $j = 1,\ldots, G$). I assume $u_{ij} > 0$ for all i, j.

The collection of all $\bar{x}_{ij}$ will be called the *initial allocation* or *initial point* and symbolized by $\bar{x}$. The collection of all x_{ij}, symbolized by x, will be called the *final allocation* or *final point*. I shall, furthermore, use x^*_{ij} and x^* to symbolize a Pareto optimal allocation.

An allocation which is feasible for the economy as a whole must satisfy

$$(1) \quad N_1 x_{i1} + \ldots + N_G x_{iG} = N_1 \bar{x}_{i1} + \ldots + N_G \bar{x}_{iG} \qquad (i = 1,\ldots, M)$$

IV. Pareto Optimal Allocations and Competitive Equilibria

As already pointed out it follows from the definition of the core that a point which is not Pareto optimal cannot belong to the core. Hence, we need only consider Pareto optimal points as candidates for belonging to the core. Furthermore we shall consider as candidates only Pareto optimal points where individuals of the same type receive the same amounts of the various goods. This implies some loss in generality, but hardly serious for our purpose. Indeed, if $N_1, N_2,\ldots, N_G$ have a greatest common divisor which is greater than one, then a very simple argument given by Jerry R. Green, which need not be repeated here, shows that core allocations have this "equal treatment property." (Convexity of preferences, as assumed above, is used in establishing this result.)

For our calculus approach it is assumed

that the optimizations defining Pareto optimal points yield interior solutions. Pareto optimal points with equal treatment as just described can then be characterized by the following conditions:

$$(2)\quad \frac{u_{1j}}{\lambda_1} = \ldots = \frac{u_{Mj}}{\lambda_M} = \mu_j \quad (j = 1, \ldots, G)$$

Pareto optimal allocations are allocations which satisfy these conditions in addition to the balances (1).

The symbols μ_j in (2), one for each type $1, \ldots, G$, are introduced for convenience as the common value of the proportions to the left. The factors of proportionality $\lambda_1, \ldots, \lambda_M$ in formula (2) can, of course, be interpreted as prices, but they are used here only as coefficients to characterize a Pareto optimal allocation, not to describe any particular institutional arrangement. Let x_{ij}^* denote the quantities corresponding to some Pareto optimal allocation, that is, an allocation satisfying (1) and (2).

I now introduce *imputed wealth*. For an individual of type j the imputed wealth in an arbitrary allocation x is defined by

$$(3)\quad y_j = \lambda_1 x_{1j} + \ldots + \lambda_M x_{Mj} \quad (j = 1, \ldots, G)$$

where the factors of proportionality are used from (2). For the initial allocation and for the Pareto optimal allocation we have, in particular, imputed wealth $\bar{y}_j$ and y_j^* respectively, defined by

$$(4)\quad \bar{y}_j = \lambda_1 \bar{x}_{1j} + \ldots + \lambda_M \bar{x}_{Mj} \quad (j = 1, \ldots, G)$$

$$(5)\quad y_j^* = \lambda_1 x_{1j}^* + \ldots + \lambda_M x_{Mj}^* \quad (j = 1, \ldots, G)$$

For the Pareto optimal allocation considered we do not necessarily have $y_j^* = \bar{y}_j$. If $y_j^* = \bar{y}_j$, then x^* represents a competitive equilibrium with prices $\lambda_1, \ldots, \lambda_M$ since the relations (2) then signify the adaptation of the various individuals to these prices and $y_j^* = \bar{y}_j$ represents the budget balance of an individual of type j. If $y_j^* \neq \bar{y}_j$ for some j, then we have a Pareto optimal point which is not a competitive equilibrium. It is well known that we may have more than one competitive equilibrium, that is, a set of equilibria. This does not matter for the following arguments.

V. The Blocking of Pareto Optimal Allocations which are not Competitive Equilibria

I now raise the question as to whether a Pareto optimal point which is not necessarily a competitive equilibrium can be blocked by any coalition. Let a possible coalition consist of $n_1, n_2, \ldots, n_G$ individuals of the various types. This coalition can, on the basis of its own initial quantities, reach any final point x which satisfies the balance relations

$$(6)\quad n_1 x_{i1} + \ldots + n_G x_{iG} = n_1 \bar{x}_{i1} + \ldots + n_G \bar{x}_{iG} \quad (i = 1, \ldots, M)$$

The imputed wealth for individuals of type j in such a point is then given by (3).

The question now is whether there exists any feasible final point x for the coalition which is considered by all members to be better than the given Pareto optimal point x^*. It follows from what has been explained that, in order to show that x^* does not belong to the core, it is sufficient to construct *one* such coalition for which *one* such point exists. We then look for a simple way to do this, not for the most general characterization of the possibilities of blocking. If we tentatively limit attention to points x which are in the neighborhood of x^*, then imputed wealth can be used as a criterion to compare x and x^*. Since we have, approximately,

$$(7)\quad U_j(x_{1j}, \ldots, x_{Mj}) - U_j(x_{1j}^*, \ldots, x_{Mj}^*) \approx u_{1j} \cdot (x_{1j} - x_{1j}^*) + \ldots + u_{Mj} \cdot (x_{Mj} - x_{Mj}^*) = [\lambda_1 (x_{1j} - x_{1j}^*) + \ldots + \lambda_M (x_{Mj} - x_{Mj}^*)]\mu_j = (y_j - y_j^*)\mu_j$$

and since $\mu_j > 0$, we have for x in the neighborhood of x^*:

$$(8)\quad y_j > y_j^* \Longrightarrow \text{an individual of type } j \text{ is better off in } x \text{ than in } x^* \quad (j = 1, \ldots, G)$$

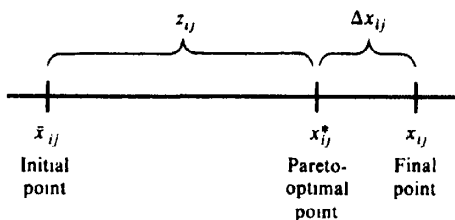



FIGURE 1

The question can now be posed as follows: Can we make $y_j > y_j^*$ hold for all j when the final point x is constrained by (6)? Introduce the following terms

$$\Delta x_{ij} = x_{ij} - x^*_{ij} \tag{9}$$

$$z_{ij} = x^*_{ij} - \bar{x}_{ij} \tag{10}$$

that is, Δx_{ij} is the deviation between the final point and the Pareto optimal point we are testing for possible blocking, and z_{ij} is the deviation between the Pareto optimal point and the initial point, as suggested in Figure 1.

We will now see if a change from x^* to x which, for each type, changes the quantities proportionately with z_{ij} will do for the purpose of blocking x^*, that is, for producing a final point which all members of the coalition find superior to x^*. We may think of this in the commodity space as drawing a straight line between the initial point $\bar{x}$ and the Pareto optimal point x^*, and then moving the final point for each group away from x^* along this ray, either towards $\bar{x}$ or further away from $\bar{x}$. Introduce the ratio

$$s_j = \Delta x_{ij}/z_{ij} \qquad (i = 1,\ldots,M;\ j = 1,\ldots,G) \tag{11}$$

If $s_j > 0$, then individuals of type j are moved further away than x^* from the initial point; if $s_j < 0$, then they are moved some distance back towards $\bar{x}$. (We may have $z_{ij} > 0$ or $z_{ij} < 0$. If, by coincidence, $z_{ij} = 0$ for some i, then also $\Delta x_{ij} = 0$, and s_j takes the value suitable for the changes in the quantities of the other commodities. If, for some j, we should happen to have $z_{ij} = 0$ for all i, then s_j is arbitrary. In the explanations which follow I shall, for brevity, neglect this special case.)

If such moves are feasible for the coalition considered, that is, satisfy (6), then we must have

$$n_1(x_{i1} - \bar{x}_{i1}) + \ldots + n_G(x_{iG} - \bar{x}_{iG}) = 0 \qquad (i = 1,\ldots,M)$$

which by use of (9)–(11) can be written as

$$n_1(1 + s_1)z_{i1} + \ldots + n_G(1 + s_G)z_{iG} = 0 \qquad (i = 1,\ldots,M) \tag{12}$$

According to (8) individuals of type j are better off at x than at x^* if we have

$$y_j - y_j^* = \lambda_1(x_{1j} - x^*_{1j}) + \ldots + \lambda_M(x_{Mj} - x^*_{Mj}) > 0$$

or, in view of (9) and (11),

$$y_j - y_j^* = (\lambda_1 z_{1j} + \ldots + \lambda_M z_{Mj})s_j > 0 \tag{13}$$

Using (4), (5), and (10), this can also be written as

$$y_j - y_j^* = (y_j^* - \bar{y}_j)s_j > 0 \tag{14}$$

This requirement determines the sign of s_j for each type j. For members of the coalition belonging to each type j we must have

$$s_j \gtreqless 0 \quad \text{according as} \quad y_j^* \gtreqless \bar{y}_j \tag{15}$$

This condition means that members of the coalition who have a larger imputed wealth at the Pareto optimal point considered than at the initial point should be moved further away from $\bar{x}$ through x^*, whereas members with higher imputed wealth in the initial situation than in the Pareto optimal point considered should be moved somewhat back from x^* towards $\bar{x}$.

I have not yet said anything about possible members for whom $y_j^* = \bar{y}_j$. This is a special case which will be disposed of later. For the moment it is assumed

$$\bar{y}_j \neq y_j^* \text{ for } j = 1,\ldots,G \tag{16}$$

We have now considered feasibility and a criterion for positive gain by members of the coalition of the various types. The feasibility condition is dependent upon the number of members of the coalition belonging to each type, i.e., on $n_1,\ldots,n_G$. The crucial question now is whether it is possible to

compose the coalition, that is, determine the numbers $n_1, \ldots, n_G$, in such a way that the feasibility condition (12) is fulfilled, while at the same time the condition (15) for a gain by all members in comparison with x^* is fulfilled.

For studying this it is convenient to introduce the proportions ν_j which members of each type form in the total coalition, i.e.,

$$(17)\quad \nu_j = \frac{n_j}{n_1 + \ldots + n_G} = \frac{n_j}{n} \qquad (j = 1, \ldots, G)$$

In terms of these proportions the feasibility requirement (12) can be written as

$$(18)\quad \nu_1(1 + s_1)z_{i1} + \ldots + \nu_G(1 + s_G)z_{iG} = 0 \qquad (i = 1, \ldots, M)$$

Observe that this condition is fulfilled for

$$(19)\quad \nu_1 = N_1/N, \ldots, \nu_G = N_G/N$$
$$s_1 = \ldots = s_G = 0$$

where $N_1, \ldots, N_G$ are the total number of individuals of each type, and N is the total number of individuals, i.e., $N = N_1 + \ldots + N_G$. This follows from the fact that the Pareto optimal point x^* must be feasible for the exchange economy as a whole, that is, we must have

$$(20)\quad N_1 x^*_{i1} + \ldots + N_G x^*_{iG} = N_1 \bar{x}_{i1} + \ldots + N_G \bar{x}_{iG} \qquad (i = 1, \ldots, M)$$

which is the feasibility condition (1) applied to the Pareto optimal point considered.

The statement just made simply means that a coalition with a composition proportional to the composition in the complete set of individuals can reach the Pareto optimal point under consideration on the basis of its own initial amounts. In order to construct a coalition which blocks the Pareto optimal point considered we shall try to find an allocation in the neighborhood of x^* which all members of the coalition prefer to x^*. We must then alter the composition of the coalition somewhat, but shall keep it *approximately* similar to the composition given by the first line of (19).

Now, in order that all members of the coalition gain by a move away from x^*, we must make $s_1, \ldots, s_G$ different from zero according to the sign pattern determined by (15). In order not to do violence to the local nature of the criterion that we use, we let $s_1, \ldots, s_G$ deviate only a little from zero. Let us for the moment treat $\nu_1, \ldots, \nu_G$ as free variables in the neighborhood of the values given by (19), restricted only by $\Sigma \nu_j = 1$. (This is a crucial point to which I shall return.) Then, for any given set of values for $s_1, \ldots, s_G$, some positive and some negative according to (15), we can clearly satisfy all equations in (18) by simply setting

$$(21)\quad \nu_1 = \frac{\alpha}{1 + s_1}\frac{N_1}{N}, \ldots, \nu_G = \frac{\alpha}{1 + s_G}\frac{N_G}{N}$$

since (18) by this insertion reduces to (20), which is known to be fulfilled. Here α is a parameter which is adjusted to that $\Sigma \nu_j = 1$.

By the procedure outlined above we have succeeded in constructing a coalition together with a feasible final point for the coalition which is superior to the Pareto optimal point x^* for all members of the coalition. By the definition of the core, we can accordingly conclude that x^* does not belong to the core. The argument is, however, not yet quite complete because of a couple of points which were temporarily put off in the development of the idea given above. Let us now return to these points.

VI. Some Special Points Needed to Complete the Argument

Let us first consider the assumption made by (16). If the Pareto optimal point considered should be such that for some type, $\bar{y}_j = y^*_j$, then individuals of this type cannot gain anything by being moved away from x^* in any direction according to the construction used above. However, it may be necessary to include a suitable number of individuals of this type in the coalition in order to give it the desired composition. For these members we set $s_j = 0$. We will then have $\nu_j = \alpha N_j/N$ according to (21). We now need these as members of the coalition, but they do not gain anything by it as compared with the Pareto optimal

point x^*. However, since other members of the coalition, for whom $\bar{y}_j \neq y_j^*$, make a strictly positive gain, then a slight transfer so as to make these special members gain also could always be carried out if this is necessary for involving them in the coalition.

As already mentioned before the case in which $\bar{y}_j = y_j^*$ for *all* $j = 1, \ldots, G$ is the case in which x^* is the special Pareto optimal point representing the competitive equilibrium, or one of these if the competitive equilibrium is not unique. In this case the procedure outlined above will not succeed in constructing another feasible final point for a coalition which is superior to x^* for all members of the coalition. This is as it should be. It is well known that the competitive equilibrium belongs to the core so that no coalition can be constructed which can block such a point. (The fact that the competitive equilibrium belongs to the core is proved by elementary methods in many expositions and will not be taken up for further consideration here.)

In connection with the comparison between y_j and y_j^*, the following point may be observed. Consider equation (14). The difference $y_j^* - \bar{y}_j$ here decides the sign of s_j. If we multiply these differences by the number of individuals of each type and add over types we get

$$(22) \quad \sum_{j=1}^{G} N_j(y_j^* - \bar{y}_j) = \sum_{i=1}^{M} \lambda_i \left(\sum_{j=1}^{G} N_j x_{ij}^* - \sum_{j=1}^{G} N_j \bar{x}_{ij}\right) = 0$$

The last equality here follows from (1) which must hold for x^*. Since all $N_j > 0$ it follows from this that when some $y_j^* > \bar{y}_j$, then there must be at least one j for which the opposite inequality holds, and vice versa. Thus, if not all $y_j^* = \bar{y}_j$, then there will be at least one j for which $s_j > 0$, and at least one j for which $s_j < 0$.

The second point which must be taken up refers to the assumption temporarily made that we could consider $\nu_1, \ldots, \nu_G$, i.e., the proportions of the representation of each type in the coalition considered, as free variables (restricted only by nonnegativity and $\Sigma \nu_j = 1$). When N is a finite integer and also $N_1, \ldots, N_G$ are integers, then we are in fact not entirely free in determining $\nu_1, \ldots, \nu_G$. These variables are defined by (17), and $n_1, \ldots, n_G$ must also be integers and restricted by $0 \leq n_j \leq N_j$. Suppose that we have tentatively determined $s_1, \ldots, s_G$ with correct signs and sufficiently small so as not to invalidate the application of our local criterion for gains. Then there may be no integers satisfying $0 \leq n_j \leq N_j$ which used in (17) produce the required $\nu_1, \ldots, \nu_G$ according to (21). The natural idea then is to make some small adjustments in $s_1, \ldots, s_G$ (without altering their signs) so as to make (21) hold good with values of $\nu_1, \ldots, \nu_G$ which can be produced by (17) with permissible integers for $n_1, \ldots, n_G$.

Now, this may be impossible if $N_1, \ldots, N_G$ are small integers. However, if $N_1, \ldots, N_G$ are large, then we are much more free in choosing $n_1, \ldots, n_G$, and it is easier to produce proportions $\nu_1, \ldots, \nu_G$ which satisfy the requirements needed for some sufficiently small $s_1, \ldots, s_G$ with correct signs. (According to what was said above in connection with (22), some of the types will be "overrepresented" in the coalition in the sense that $\nu_j > N_j/N$, and some types will be "underrepresented" in the sense that $\nu_j < N_j/N$. The factor α used to secure $\Sigma \nu_j = 1$ will be near to unity when $s_1, \ldots, s_G$ are small.)

If we increase $N_1, \ldots, N_G$ beyond all limits, then we approach a situation in which the restriction that $n_1, \ldots, n_G$ have to be integers is no longer an effective restriction on the possibilities for choosing $\nu_1, \ldots, \nu_G$. Then the construction of the coalitions as given above can be carried out for any Pareto optimal point x^* which is not a competitive equilibrium, that is, for any x^* for which at least one type (and then necessarily at least two) have $y_j^* \neq \bar{y}_j$. This shows that when the number of individuals of all types increases beyond all limits, then only competitive equilibrium solutions remain in the core. (In order to make the comparison between smaller and larger economies meaningful, it is easiest to think of the larger economy as one in which the

number of individuals of each type has been blown up proportionately. Then we may speak about "the same point" x^* in the smaller and the larger economy, the only difference being in the absolute numbers of individuals enjoying the various commodity bundles.)

VII. A Final Remark

The construction presented above can also be used to say something more intuitive about the size of the core when the number of individuals is finite, and in general other (Pareto optimal) points besides the competitive equilibrium belong to the core. For instance, if the indifference surfaces corresponding to the utility functions of individuals of the various types are very strongly curved, then there will be less freedom in the choice of $s_1, \ldots, s_G$, while smaller curvature makes for wider ranges of permissible choices of $s_1, \ldots, s_G$. The less free we are in choosing $s_1, \ldots, s_G$, the more difficult will it be to find permissible $\nu_1, \ldots, \nu_G$ when we have a limited number of individuals of each type to select $n_1, \ldots, n_G$ from. Thus, for an economy with a given number of individuals, the blocking procedure used here seems to be more powerful in excluding points from the core when there is a moderate curvature than when there is strong curvature in the indifference surfaces in the neighborhood of the point tested. By similar reasoning one may also get the impression that it will normally be easier to exclude points which are far away from the competitive equilibrium than points in its neighborhood. However, these suggestions are only hints about directions in which the arguments can be developed. A complete analysis of the question as to which points belong to the core and which ones do not, for a given number of individuals of each type, is a much more difficult task than the one tackled above. In order to show that a point does *not* belong to the core, it is sufficient to construct *one* particular coalition which is able to block the point in *one* particular way as we have done above. In order to show that a point belongs to the core, one must show that all possible coalitions with all their feasible reallocations fail to produce a point which is superior to the point considered for all members. Except for competitive equilibrium points, this is usually a complicated matter.

REFERENCES

G. Debreu and H. Scarf, "A Limit Theorem on the Core of an Economy," *Int. Econ. Rev.*, Sept. 1963, *4*, 235–46.

——— **and** ———, "The Limit of the Core of an Economy," in C. B. McGuire and Roy Radner, eds., *Decision and Organization: A Volume in Honor of Jacob Marschak*, Amsterdam 1972.

J. R. Green, "On the Inequitable Nature of Core Allocations," *J. Econ. Theory*, Apr. 1972, *4*, 132–43.

Edmond Malinvaud, *Lectures on Microeconomic Theory*, Amsterdam 1972.

Peter Newman, *The Theory of Exchange*, Englewood Cliffs 1965.

L. S. Shapley and M. Shubik, "Concepts and Theories of Pure Competition," in Martin Shubik, ed., *Essays in Mathematical Economics in Honor of O. Morgenstern*, Princeton 1967.

M. Shubik, "Edgeworth Market Games," in A. W. Tucker and Robert D. Luce, eds., *Contributions to the Theory of Games*, *Annals of Mathematical Studies*, Vol. IV, Princeton 1959.

Inflation, Hedging, and the Demand for Money

By C. F. J. BOONEKAMP*

Casual observation reveals that the future rate of inflation is not known with certainty. Money, consequently, has an uncertain command over future consumption. Explanations of the portfolio demand for money have traditionally assumed as have, indeed, most models of portfolio behavior, that the decision maker selects a portfolio so as to maximize the expected utility of terminal wealth which is, implicitly or otherwise, assumed to be defined in nominal terms. Economic theory, however, postulates that real rather than nominal factors condition economic motivation. Theory would thus require that we be concerned with the expected utility of real wealth in the determination of portfolio decisions. If it is assumed that the price level is known with certainty, then clearly a one-to-one relationship exists between real and nominal wealth and it is then of no consequence whether the portfolio problem is set out in real or nominal terms. This approach, in that it presupposes a zero variance in the rate of inflation, assumes the absence of purchasing-power risk. Our current economic climate does not permit such an assumption. Failure to satisfy the zero variance assumption implies that the relationship between real and nominal wealth is uncertain. Consequently we cannot assert that the only difference between the maximization of attainable utility from real and nominal wealth is one of scaling. The presence of purchasing-power risk might then be expected in general to influence the portfolio demand for money and, in fact, for all assets.

In this paper I discuss the implications of purchasing-power risk for the portfolio demand for money, in particular the consequences of price uncertainty for the allocation of a household's wealth between two monetary assets, one of which is money. It is demonstrated that the extent to which an asset protects against depreciation in the real value of a unit of nominal balances is of importance to the portfolio chosen. The portfolio demand for an asset is consequently a function not only of the well-known Tobin-Markowitz speculative motive but also a hedging motive. This follows directly from the fact that, in general, there is no riskless asset and therefore no riskless portfolio in real terms. Nevertheless the hedging properties of an asset may serve to reduce the variability of future real wealth. The Tobin-Markowitz solution is then a special case of the solution derived when price uncertainty is explicitly considered. The analysis also provides a generalization of Kenneth Arrow's conjecture that the wealth elasticity of demand for money is at least one if the index of relative risk aversion is an increasing function of initial wealth.

The fact that money has an uncertain command over future consumption has only recently been formally incorporated in the literature.[1] Fischer Black in an important paper implicitly deals with price uncertainty by considering a capital-asset pricing model in the absence of a risk-free asset; he does not however identify a hedging motive, nor does he relate price uncertainty to the portfolio demand for money. Richard Roll formally stated the problem; he derived first-order conditions for a solution but, as a result of the very general probability distributions over prices which he allows, he was not explicit about

*University of British Columbia. I am grateful to G. C. Archibald, C. Azariadis, G. H. Borts, J. G. Cragg, R. Davidson, E. Diewert, J. Hanson, R. A. Jones, K. Nagatani, D. Rose, E. Schwartz, and W. Schworm for many helpful comments and suggestions. Any remaining errors are of course my responsibility.

[1] Its possible significance has, however, been recognized for some time (see for example James Tobin 1958, 1965).

the portfolio chosen. Nahum Biger conducts an empirical investigation of inflation assessment and portfolio selection; he is not, and for his purposes need not be, explicit about the portfolio chosen. Stanley Fischer examined the demand for index bonds; within the context of a continuous-time model of consumption-saving and portfolio-allocation decisions, and under the assumption of small price risks, he partially characterized the portfolio chosen in terms of the first two moments of the real rates of return of the assets. Robert A. Jones examined the link between the use of money as a medium of exchange and as a standard of deferred payment. As a first step in the analysis, using results from duality theory and under the assumption of small price risks, he established a general relationship between an individual's consumption preferences and a desired portfolio of claims to future consumption. I too shall assume that price risks are small, will use a simple result from duality theory, and will consequently be able to derive an explicit and potentially testable solution for the portfolio chosen.

This paper is divided into four sections. Section I will state the objective of the household and the assumptions under which the analysis is pursued. Section II will be devoted to the determination of the real rates of return to the assets and to the riskiness of those rates of return. Section III will solve the portfolio-choice problem and discuss the implications of the result. Section IV presents the conclusions of the paper. A glossary of notation is presented in Appendix A. An apparently anomalous result, that an increase in the expected opportunity cost of holding money might lead to an increase in desired nominal balances, is discussed in Appendix B.

I

It is assumed that households derive utility only from consumption whence they are concerned with the real, rather than nominal, value of their wealth. Given wealth, the household selects a portfolio so as to maximize the expected utility from the consumption of goods to be purchased with the portfolio in the future. Cast in a two-period framework the households portfolio budget constraint is[2]

$$W^o = M + P_B^o B$$

where W^o = nominal wealth at the end of the first period
M = nominal balances
B = bonds[3]
P_B^o = period-one price of a bond

The constraint does not preclude "long" or "short" positions on either of the assets. The household is able to define a composite good, X.[4] The utility function $U = u(X)$ is assumed to be at least twice differentiable and concave throughout its domain. The period-one price of X is P_X^o. The prices of X and B that will prevail in the next period are unknown to the household in the first period; the household has, however, subjective beliefs about these prices which are characterized by the subjective means of period-two prices:[5]

$$\bar{P}_X = E[P_X] \gtreqless P_X^o$$
$$\bar{P}_B = E[P_B] \gtreqless P_B^o$$

and the subjective covariance matrix of period-two prices:

[2]The two-period model does introduce the fiction of a fixed investment period; the suggestion is, therefore, that decision and transactions costs are of sufficient size to warrant a certain inertia in portfolio composition. This is not totally satisfactory for inasmuch as these costs affect the level of expected wealth, and thus the index of relative risk aversion, they will influence the relative share of each asset in the portfolio. The length of the period may, however, be defined by the size of the costs and can consequently be made small if the costs are small relative to portfolio size; moreover, the two-period model facilitates emphasis on the importance of uncertainty with respect to the price level.

[3]A bond is to be regarded as a generic term for any financial asset other than money.

[4]A composite good and the price index are well defined if the household has, as is assumed, homothetic preferences over goods.

[5]$\bar{P}_B$ includes the accrued interest, if any, due to the asset.

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XB} \\ \sigma_{BX} & \sigma_B^2 \end{bmatrix}$$

The dollar acts as the unit of account and its nominal price is always equal to one. It is assumed that price risks are small, by which is meant that period-two price levels will lie in some small neighborhood about their present levels with probability one.

The household at the end of the first period allocates its wealth amongst dollars and bonds so as to maximize the expected utility from the consumption of the composite good to be purchased with the portfolio at the end of the next period. Its objective is, therefore, to

$$(1) \quad \max_{B,M,X} \; E_{P_X,P_B} \, [u(X)]$$

subject to: $W^o = M + P_B^o B$

$$P_X X = M + P_B B$$

where $M + P_B B = W$ is the uncertain nominal value of the portfolio. It is clear that the dollar is a safe asset in nominal terms but owing to composite-good price uncertainty it is not a safe asset in real terms. Bonds are risky in both nominal and real terms. The household, explicitly concerned with the real value of its portfolio, is thus denied the recourse of a safe asset in selecting its portfolio.[6]

The problem, as stated in (1), represents a departure from traditional portfolio demand for money models. Were it assumed that future utility depended solely on the future *nominal* value of the portfolio then we should have employed the Tobin-Markowitz model: a risk-averse household would allocate its wealth on the basis of the differing expected nominal rates of return, and riskiness of return, to dollars and bonds. Dollars would be a safe asset. The present formulation, however, explicitly recognizes that future utility is dependent *not* on the level of nominal wealth available from the portfolio, but rather on the quantity of X which can actually be consumed; if there is then uncertainty about the future price of the composite good there is not a one-to-one relationship between future wealth and attainable utility. The household, in selecting its portfolio, must then direct its attention to the expected real rates of return and the riskiness of those rates of return. The impact of an uncertain rate of inflation on the expected yields to the assets proves to be a distinguishing feature between the Tobin-Markowitz analysis and that presented here. It is to a discussion of these yields and their riskiness that I now turn.

II

Within the discrete framework adopted the second-period rate of inflation in the price of X of given by

$$(2) \qquad \Pi_X = \frac{P_X - P_X^o}{P_X^o}$$

where Π_X is a random variable, the (subjective) probability distribution of which is characterized by the expected rate of inflation

$$\bar{\Pi}_X = E[\Pi_X]$$

and the variance:

$$V[\Pi_X] = \frac{\sigma_X^2}{(P_X^o)^2} = \sigma_{\Pi X}^2$$

The real rate of return on a dollar is

$$(3) \qquad r_M = \frac{P_X^o}{P_X} - 1 = \frac{1}{1 + \Pi_X} - 1$$

where r_M is a random variable. Since the function $1/(1 + \Pi_X) - 1$ is, over the relevant range,[7] a convex function of Π_X by combining (3) with Jensen's inequality it follows that

[6]The household is precluded from buying X in the first period and then storing it for use in the second period.

[7]The first and second derivative of $1/(1 + \Pi_X) - 1$ with respect to Π_X are respectively: $-1/(1 + \Pi_X)^2 < 0$ and $2/(1 + \Pi_X)^3 > 0$ iff $\Pi_X > -1$. If $\Pi_X \leq -1$ it implies that $P_X \leq 0$, a prospect to which I assign zero probability.

$$(4)\quad E[r_M] = E\left[\frac{1}{1+\Pi_X} - 1\right] \geq \frac{1}{1+E[\Pi_X]} - 1$$

That is, the expected real rate of return to a unit of nominal balances is, given a constant mean, greater when the rate of inflation is random than when the next period rate of inflation is known with certainty.[8] Under the assumption of small price risks (4) is adequately approximated by a Maclaurin series in which all terms beyond those associated with the second derivative are ignored. Thus

$$(5)\quad E[r_M] = E[-\Pi_X + \Pi_X^2 - \ldots] \geq -E[\Pi_X] + E[\Pi_X]^2 - \ldots$$

After taking the expectation (5) yields

$$(6)\quad E[r_M] = \bar{r}_M = -\bar{\Pi}_X + \sigma_{\Pi X}^2 + \bar{\Pi}_X^2 \geq -\bar{\Pi}_X + \bar{\Pi}_X^2$$

The square of the expected rate of inflation appears on both sides of the inequality because even when the future rate of inflation is known with certainty, $-\bar{\Pi}_X$ is only a first-order approximation to the expected real rate of return. If, as has been assumed, $\bar{\Pi}_X$ is small, $\bar{\Pi}_X^2$ is safely ignored.[9] Thus (6) may be rewritten as

$$(7)\quad E[r_M] = \bar{r}_M = -\bar{\Pi}_X + \sigma_{\Pi X}^2 \geq -\bar{\Pi}_X$$

and the expected real rate of return to holding a dollar is

$$(8)\quad E[r_M] = \bar{r}_M = -\bar{\Pi}_X + \sigma_{\Pi X}^2$$

Equation (8) shows clearly that $\bar{r}_M$ is *not*, as it is often specified to be, the negative of the expected rate of inflation; that specification would be correct were there no uncertainty about the future rate of inflation, i.e., if $\sigma_{\Pi X}^2 = V[\Pi_X] = 0$. Equation (8) shows, too, that $\bar{r}_M$ increases with the variance, holding the mean constant, and decreases with mean, holding the variance constant. Had the Maclaurin series used to derive (8) been further expanded it would have shown that $\bar{r}_M$ increases with the even moments and decreases with the odd moments. If it is thought that the range of $\bar{\Pi}_X$ is large it may be necessary to take account of these higher order moments.

Using the same technique as employed in the determination of (8) yields, as the variance of the real rate of return on a dollar,

$$(9)\quad V[r_M] = \sigma_X^2/(P_X^o)^2 = \sigma_{\Pi X}^2$$

Following Tobin (1965), $V[r_M]$ serves as a measure of the riskiness of holding a dollar.

The distribution of the real rate of return on a dollar is therefore described as[10]

$$r_M \sim (-\bar{\Pi}_X + \sigma_{\Pi X}^2, \sigma_{\Pi X}^2)$$

The real rate of return to holding a dollar's worth of bonds is the additional amount of X that can be purchased in the next period compared to the amount of X that can now be purchased. Thus

$$(10)\quad r_B = \frac{P_B/P_X}{P_B^o/P_X^o} - 1$$

where r_B is a random variable the distribution of which is defined by the distributions of P_X and P_B. Expanding (10) in a Taylor series about P_B^o, P_X^o and taking expectations of both sides yields

$$(11)\quad E[r_B] = \bar{r}_B = \bar{\Pi}_B - \bar{\Pi}_X + \sigma_X^2/(P_X^o)^2 - \frac{\sigma_{XB}}{P_X^o P_B^o} = \bar{\Pi}_B - \bar{\Pi}_X + \sigma_{\Pi X}^2 - \sigma_{\Pi XB}$$

In (11), $\bar{r}_B$ is the expected real rate of return to a dollar's worth of bonds, $\bar{\Pi}_B$ is the ex-

[8]The author and David Donaldson presented a rigorous explanation of this result. It has also been noted by Benjamin Eden and by Fischer.

[9]If the future rate of inflation is known with certainty, i.e., if point anticipations are held, then

$$\bar{r}_M = E[r_M] = \frac{1}{1-\bar{\Pi}_X} - 1 = -\bar{\Pi}_X + \bar{\Pi}_X^2 - \bar{\Pi}_X^3 + \bar{\Pi}_X^4 - \ldots$$

where $\bar{\Pi}_X^2$ and beyond are normally ignored; during periods of hyperinflation the ignored terms are, however, of conceivable importance.

[10]This result is akin to that of Fischer, with the difference only that in Fischer's case the variance, as a result of describing the price dynamics in a continuous-time framework by an Ito process, goes to infinity as the time horizon goes to infinity.

pected nominal yield on bonds,[11] and $\sigma_{XB}/P_X^o P_B^o = \sigma_{\Pi XB}$ is the covariance between the rate of inflation in the price of X and the nominal yield on bonds. The covariance term might be expected to enter (11) with a positive sign. Since the dollar is the medium of exchange, the chain of conversion from bonds to X runs through dollars whence $\bar{r}_B$ is like the union of two non-mutually exclusive sets. The argument implies that the covariance—the intersection of the two sets r_B, r_M—enters with a negative sign.

Several interesting points emerge from (11). First, the expected real rate of return on bonds is equal to the expected nominal rate $\overline{\Pi}_B$, minus the expected rate of inflation only in the absence of composite-good price uncertainty, i.e., if $\sigma^2_{\Pi X} = 0$, which implies, of course, that $\sigma_{\Pi XB}$ is zero. Second if P_B and P_X are independent, such that the covariance term is zero, then the expected net real return to holding bonds over money, $\bar{r}_B - \bar{r}_M = \overline{\Pi}_B$, is simply the expected nominal rate of return to bonds; bonds are then neither a positive nor a negative *hedge* against deflation in the value of the dollar. The extent to which bonds are a hedge will, as shown in Section III, be of considerable concern to the household. Third, and this applies equally to $\bar{r}_M$, even if the expected nominal rate is zero then, in the absence of price certainty, the expected real rate is *not* zero; this fact is of importance to portfolio determination.

Using again the technique employed above yields for the variance of the real rate of return on bonds

$$(12)\quad V[r_B] = \sigma_X^2/(P_X^o)^2 + \sigma_B^2/(P_B^o)^2 - \frac{2\sigma_{XB}}{P_X^o P_B^o} = \sigma^2_{\Pi X} + \sigma^2_{\Pi B} - 2\sigma_{\Pi XB}$$

where $\sigma^2_{\Pi B}$ is the variance of the nominal yield on bonds. Note that (12), like (11), reflects the fact the the dollar is the medium of exchange. Note too that the interdependence between the two price levels serves to influence the real risk of holding bonds, an intuitively tenable result.

The distribution of the real rate of return to a dollar's worth of bonds is therefore described as

$$r_B \sim (\overline{\Pi}_B - \overline{\Pi}_X + \sigma^2_{\Pi X} - \sigma_{\Pi XB},\ \sigma^2_{\Pi X} + \sigma^2_{\Pi B} - 2\sigma_{\Pi XB})$$

[11] $\Pi_B = (P_B - P_B^o)/P_B^o$, a random variable whose distribution is characterized by $\overline{\Pi}_B = E[\Pi_B]$ the subjective mean of the yield, and $\sigma_B^2/(P_B^o)^2$, the subjective variance.

III

In this section the household's portfolio choice problem is solved and discussed. As shown in expression (1) the household is to maximize expected utility. In arriving at a portfolio decision the household must, in a dynamic programming fashion, take the second-period consumption decision into account. Formally, once P_X and W have been revealed the consumption choice problem is solved by

$$(13)\quad \max\ L(X,\lambda) = u(X) + \lambda[W - XP_X]$$

where λ is the Lagrangian multiplier. This yields, trivially, that

$$(14)\qquad X^d = g(W, P_X) = \frac{W}{P_X}$$

where X^d is the demand for good X. Given W and P_X, (14) allows attainable utility in period two to be expressed as

$$(15)\quad U = H(W, P_X) = u[g(W, P_X)]$$

The household's objective is therefore rewritten as

$$(16)\qquad \max_{B,M}\ \underset{P_X, P_B}{E}\ [H(W, P_X)]$$

subject to $W^o = M + P_B^o B$.

The expectation of a Taylor series expansion of $H(W, P_X)$ about W^o, P_X^o leads to an adequate approximation of $E[H(W, P_X)]$. Thus,[12]

[12] A second-order Taylor expansion is used and thus a mean-variance approximation to the expected utility hypothesis is adopted. Paul Samuelson has, for the case when the risks associated with the assets are small, provided a formal justification for this approximation technique. The assumption of small price risks places the analysis in this paper within the parameters of the justification.

$$(16') \quad E[H(W,P_X)] = H(W^o,P_X^o) + H_1(\bar{P}_B - P_B^o)B + H_2(\bar{P}_X - P_X^o) + 1/2H_{11}\sigma_B^2 B^2 + 1/2H_{11}(\bar{P}_B - P_B^o)^2 B^2 + 1/2H_{22}\sigma_X^2 + 1/2H_{22}(\bar{P}_X - P_X^o)^2 + H_{12}\sigma_{XB}B + H_{12}(\bar{P}_X - P_X^o)(\bar{P}_B - P_B^o)B$$

where subscripts on H denote partial derivatives, such that for example H_1 is the partial derivative of $H(W, P_X)$ with respect to W, evaluated at (W^o, P_X^o). The manipulations $(P_X - P_X^o) = (P_X - \bar{P}_X) + (\bar{P}_X - P_X^o)$ and $(W - W^o) = B(P_B - \bar{P}_B) + B(\bar{P}_B - P_B^o)$ have been used; the second manipulation implies the substitution of the portfolio budget constraint into the expansion of the utility function and, therefore, that only B remains as a choice variable. The first- and second-order conditions for the expected utility maximum yield

$$(17) \quad \frac{dE[H(\cdot)]}{dB} = H_1(P_B - P_B^o) + H_{11}\sigma_B^2 B + H_{11}(P_B - P_B^o)^2 B + H_{12}\sigma_{XB} + H_{12}(\bar{P}_X - P_X^o)(\bar{P}_B - P_B^o) = 0$$

$$(17') \quad \frac{d^2E[H(\cdot)]}{dB^2} = H_{11}\sigma_B^2 + H_{11}(\bar{P}_B - P_B^o)^2 < 0 \quad \text{iff } H_{11} < 0$$

If, therefore, the household is risk averse, $H_{11} < 0$, the second-order condition for an interior maximum is satisfied. $H_{11} = 0$ implies risk neutrality and the possibility of a corner solution to the portfolio problem. From (17) the portfolio demand for bonds is

$$(18) \quad B^* = \frac{-H_1(\bar{P}_B - P_B^o)}{H_{11}\{(\bar{P}_B - P_B^o)^2 + \sigma_B^2\}} - \frac{H_{12}\{\sigma_{XB} + (\bar{P}_B - P_B^o)(\bar{P}_X - P_X^o)\}}{H_{11}\{(\bar{P}_B - P_B^o)^2 + \sigma_B^2\}}$$

By Roy's Theorem[13]

$$(18') \quad H_2 = -H_1 X^d$$

In other words, the loss in attainable utility from a rise in the price of X equals the marginal utility of wealth multiplied by the demand for X. Consequently

$$(18'') \quad H_{12} = -H_{11}X^d - H_1\frac{dX^d}{dW} = -H_{11}[1 - \frac{1}{RRA}]\frac{W}{P_X}$$

where the derivatives and variables are evaluated at (W^o, P_X^1) and $RRA = -H_{11}W^o/H_1$ is the Arrow-Pratt index of relative risk aversion (RRA).[14] Substituting (18″) into (18), converting to portfolio shares and dropping those second-order terms which are negligible because of the assumption of small price risks yields

$$(19) \quad A^* = \frac{1}{RRA}\cdot\frac{\bar{\Pi}_B}{\sigma_{\Pi B}^2} + \frac{\sigma_{\Pi XB}}{\sigma_{\Pi B}^2}(1 - \frac{1}{RRA})$$

[13]A discussion of the theorem is to be found in Erwin Diewert.

[14]Equations (18′) and (18″) can be used to derive the interesting result that a household may prefer uncertainty with respect to the price of X over the price stabilized at its present level. Proceed as follows: from (16′)

$$1/2\, H_{22}\{\sigma_X^2 + (\bar{P}_X - P_X^o)^2\}$$

enters $E[H(W, P_X)]$ with a positive sign. From (18′)

$$H_{22} = -H_{12}X^d - H_1\frac{\partial X^d}{\partial P_X} = -H_{12}X^d + H_1\frac{W}{(P_X)^2}$$

which by (14) and (18″)

$$= H_{11}[1 - \frac{1}{RRA}]\frac{W^2}{(P_X)^2} + H_1\frac{W}{(P_X)^2} = H_1[1 - RRA]\frac{W}{(P_X)^2} + H_1\frac{W}{(P_X)^2} = H_1[2 - RRA]\frac{W}{(P_X)^2} \geq 0 \text{ if } RRA \leq 2$$

where, again, all derivatives and variables are evaluated at (W^o, P_X^o). Thus

$$1/2\, H_{22}\{\sigma_X^2 + (P_X - P_X^o)^2\} > 0 \text{ for } RRA < 2$$

The conclusion is, therefore, that a household with a *given* level of nominal wealth W^o, an $RRA < 2$, and homothetic preferences over goods such that X and P_X are defined will prefer consumption good price uncertainty over the price stabilized at P_X^o. This is the essence of the "Waugh paradox" which conjectures that consumers may prefer price uncertainty over stabilized prices. Jones and Giora Hanoch have noted similar results to the above.

where $A^* = B^*P_B^o/W^o$, the household's demand for bonds expressed as a fraction of its present wealth. Simple manipulation yields

$$(19') \quad A^* = \frac{1}{RRA}\frac{(\bar{r}_B - \bar{r}_M)}{\sigma_{\Pi B}^2} + \frac{\sigma_{\Pi XB}}{\sigma_{\Pi B}^2}$$

where, using (8) and (11), $\bar{r}_B - \bar{r}_M = \bar{\Pi}_B - \sigma_{\Pi XB}$, the net real yield on bonds. The term $\sigma_{\Pi XB}/\sigma_{\Pi B}^2$ warrants inspection. Let the household be infinitely risk averse, $RRA \rightarrow \infty$, such that it selects the minimum variance portfolio or, alternatively stated, such that it selects a portfolio so as to *maximize* the *minimum* level of attainable utility in the next period; it will then choose

$$A^{**} = \lim_{RRA \rightarrow \infty} A^* = \frac{\sigma_{\Pi XB}}{\sigma_{\Pi B}^2}$$

In combination with the portfolio budget constraint (19) yields the household's demand for money.

$$(19'') \quad \frac{M^*}{W^o} = 1 - \frac{B^*P_B^o}{W^o} = 1 - A^*$$

At this point it is useful to solve the household's choice problem under the traditional assumption of consumption-good price certainty. This step facilitates comparison of the above result to that derived using the Tobin-Markowitz model. The assumption of certainty with respect to the future rate of inflation allows the household's objective to be rewritten as

$$(20) \quad \max_{B,M} \underset{P_B}{E}\,[V(W)]$$

subject to $\quad W^o = M + P_B^o B$

where $V(W) = H(W, \bar{P}_X)$.

The methods used above yield

$$(20') \quad E[V(W)] = V(W^o) + V_W(\bar{P}_B - P_B^o)B + 1/2\,V_{WW}\sigma_B^2 B^2 + 1/2\,V_{WW}(\bar{P}_B - P_B^o)^2 B^2$$

Maximizing (20') with respect to B, converting to portfolio shares and recognizing $-W^o V_{WW}/V_W$ as the index of relative risk aversion yields

$$(20'') \quad N^* = \frac{1}{RRA}\frac{\bar{\Pi}_B}{\sigma_{\Pi B}^2}$$

where $N^* = B^*P_B^o/W^o$, a household's demand for bonds under the assumption of certainty about the future rate of inflation.[15] Using the portfolio budget constraint yields the demand for money:

$$(20''') \quad \frac{M^*}{W^o} = 1 - N^*$$

The parameters $\bar{\Pi}_B$ and $\sigma_{\Pi B}^2$ are, respectively, the nominal expected return and risk of return to bonds; they are the parameters of concern to a risk-averse household seeking only to maximize the expected utility from nominal wealth. The demand for money so derived is, therefore, designated as the *speculative* demand for money.

An examination of (19) shows that the portfolio demand for bonds derived under the assumption of an uncertain rate of inflation differs from the speculative demand by the term

$$\frac{\sigma_{\Pi XB}}{\sigma_{\Pi B}^2}\left(1 - \frac{1}{RRA}\right) = \frac{\rho\sigma_{\Pi X}\sigma_{\Pi B}}{\sigma_{\Pi B}^2}\left(1 - \frac{1}{RRA}\right)$$

A strictly positive correlation coefficient ρ indicates that the random variable Π_X and Π_B move in the same direction, and vice versa if ρ is negative. The covariance term is thus an indication of the extent to which bonds are a hedge against depreciation in the value of the dollar. The whole term above represents the contribution of the hedging properties of bonds to a household's demand for bonds. The term is therefore designated as the *hedging effect*.

A household whose utility is defined on goods has thus two distinct and not necessarily compatible motives for holding nominally risky assets. Traditional analysis,

[15]Equation (20″) is akin to the classic Tobin (1958) result. Had the second-order terms which are negligible because of the assumption of small price risks not been dropped (20″) would read as

$$N^* = \frac{1}{RRA}\cdot\frac{\bar{\Pi}_B}{\sigma_{\Pi B}^2 + \bar{\Pi}_B^2}$$

which is Tobin's result.

concerned only with nominal yields and risks, misses the plausible hedging motive for the inclusion of an asset in a portfolio. A simple example highlights the point. Assume that $\overline{\Pi}_B = 0$ but that $\sigma^2_{\Pi B} > 0$. Dollars and bonds are then both expected to yield a zero nominal return. Because bonds unlike dollars are a risky asset in nominal terms the traditional analysis would find no motive for portfolio diversification. The above analysis, because of the hedging effect, suggests the contrary. The result pertains because, whereas the selection of a low risk portfolio in nominal terms when the price of X is fixed—or known with certainty—corresponds to choosing low variability in future attainable utility, the selection of a low risk portfolio in nominal terms when the price of X is random corresponds to choosing higher variability in future attainable utility. The hedging properties of bonds can serve to reduce the latter variability and thus give rise to diversification.

An examination of (19) shows that there are only three circumstances under which the demand for bonds—and, hence, for money—may be safely represented by the speculative motive alone. The first case is when the future price level is known with certainty; $\sigma_{\Pi X}$ then is zero and the solution reduces to the Tobin-Markowitz portfolio. The latter portfolio is thus a special case of the solution derived when price uncertainty is explicitly considered. The second case is when $\rho = 0$; bonds are then not viewed as a hedge against inflation and the hedging effect consequently reduces to zero. The third, and interesting case, is that when the index of relative risk aversion is equal to one; the household then seemingly ignores the hedging effect and selects the portfolio dictated by the conventional analysis. The explanation for this phenomenon is to be found in the fact that for an $RRA = 1$ the indirect utility function is logarithmic in nominal wealth and prices, i.e.,

$$\lim_{RRA \to 1} H(W, P_X) = \ln W - \ln P_X$$

Thus, although attainable utility does depend on P_X, *marginal* utility from additional wealth does not. The household, consequently, maximizes expected utility by selecting a portfolio *as if* the price of X were fixed, and therefore it chooses the conventional case portfolio.

The properties of the portfolio demand for bonds—and, hence, of the portfolio demand for money—are readily derived. It is convenient for that purpose to rewrite (19) as

$$(21) \qquad A^* = \frac{1}{RRA}\frac{\overline{\Pi}_B}{\sigma^2_{\Pi B}} + \frac{\rho\sigma_{\Pi X}}{\sigma_{\Pi B}}\left(1 - \frac{1}{RRA}\right)$$

An obviously important variable in equation (21) is ρ; I therefore examine the properties of (21) for the three separate ranges of values, $\rho = 0$, $\rho > 0$, $\rho < 0$, that the correlation coefficient can take on.

Case A, $\rho = 0$: Under these circumstances, because bonds are not viewed as an effective hedge against inflation, the household's demand for bonds is represented by the speculative motive alone. The properties of (21) are then

$$(a) \qquad \frac{\partial A^*}{\partial \overline{\Pi}_B} = \frac{1}{RRA}\cdot\frac{1}{\sigma^2_{\Pi B}} > 0$$

$$(b) \qquad \frac{\partial A^*}{\partial(\sigma^2_{\Pi B})} = -\frac{1}{RRA}\cdot\frac{\overline{\Pi}_B}{(\sigma^2_{\Pi B})^2} < 0$$

$$(c) \qquad \frac{\partial A^*}{\partial(RRA)} = -\frac{1}{(RRA)^2}\cdot\frac{\overline{\Pi}_B}{\sigma^2_{\Pi B}} < 0$$

$$(d) \qquad \frac{\partial A^*}{\partial W^o} = \frac{\partial A^*}{\partial(RRA)}\cdot\frac{d(RRA)}{dW^o} \gtreqless 0$$

where $\dfrac{d(RRA)}{dW^o} \lesseqgtr 0$

Notice that if the RRA is constant then the wealth elasticity of demand for bonds (and money) is unity; if, however, the RRA is an increasing function of initial wealth then the wealth elasticity of demand for money is *at least one*. Arrow showed that "it follows from the hypothesis of increasing relative risk aversion that the wealth elasticity of demand for cash is at least one" (p. 44). His result was derived within the framework of a portfolio choice model between a risky asset and a secure asset,

cash. The result thus generalizes Arrow's result to the case of two unsafe assets and $\rho = 0$. The remainder of the above properties are well established in the portfolio literature and need no further elaboration.

Case B, $\rho < 0$: In this case bonds are a negative hedge against depreciation in the value of the dollar. Since the speculative motive is additive to the hedging effect the analysis is restricted to the hedging demand, where the latter is expressed as

$$(21') \qquad A_H^* = \frac{\rho\sigma_{\Pi X}}{\sigma_{\Pi B}}\left(1 - \frac{1}{RRA}\right)$$

It is immediately obvious that A_H^* disappears if the $RRA = 1$. The reason for this was discussed earlier and is not repeated. Note that if the household is linear in risk, $RRA = 0$, then A_H^*, given $\rho < 0$, is infinite and the household consequently adopts an infinitely "short" position on money: it does so because it ignores risk and opts exclusively for the asset, bonds, expected to yield the highest rate of return. If, on the other hand, the household is infinitely risk averse, $RRA \rightarrow \infty$, the hedging effect reduces to $\rho\sigma_{\Pi X}/\sigma_{\Pi B}$, the speculative motive reduces to zero and the household adopts a "long" position on money despite the fact, as reference to (8) and (11) shows, that the expected net real yield to bonds is positive. The economic explanation is that the household is prepared to pay a large premium to insure a basic minimum level of utility in the next period.

In general, if $\rho < 0$, for A_H^* to yield a positive contribution to the speculative demand for bonds it is necessary that the RRA be less than 1, that is, that the household be less risk averse than with a logarithmic utility function. The properties of the hedging effect under these circumstances are

$$\text{(a)} \qquad \left.\frac{\partial A_H^*}{\partial \sigma_{\Pi X}}\right|_{0<RRA<1,\rho<0} > 0$$

Thus, as dollars become riskier in real terms, there is a move into bonds.

$$\text{(b)} \qquad \left.\frac{\partial A_H^*}{\partial |\rho|}\right|_{0<RRA<1,\rho<0} > 0$$

where $|\rho|$ is the absolute value of ρ. This result derives from the fact (shown in Section II) that as ρ decreases both $\bar{r}_B$ and $V[r_B]$ increase; for an $RRA < 1$, the increase in $\bar{r}_B$ is sufficient to compensate for the increase in $V[r_B]$. This is discussed further in Appendix B.

$$\text{(c)} \qquad \left.\frac{\partial A_H^*}{\partial \sigma_{\Pi B}}\right|_{0<RRA<1,\rho<0} < 0$$

The sign on the speculative motive is the same; thus, as bonds become a riskier asset, there is a move into dollars.

$$\text{(d)} \qquad \left.\frac{\partial A_H^*}{\partial (RRA)}\right|_{0<RRA<1,\rho<0} < 0$$

Thus, as the household becomes more risk averse it moves out of the nominally risky asset, bonds, and into the nominally safe asset, dollars.

$$\text{(e)} \quad \left.\frac{\partial A_H^*}{\partial W^o}\right|_{0<RRA<1,\rho<0} = \frac{\partial A_H^*}{\partial (RRA)} \frac{d(RRA)}{dW^o} \lesseqgtr 0$$

where $d(RRA)/dW^o \lesseqgtr 0$. If, therefore, the RRA is an increasing function of W^o proportionately more of an increase in the "initial endowment" is devoted to the nominally safe asset, money, then to the nominally risky asset, bonds. This property, in fact, relies only on the RRA being positive. Thus, again, if the RRA is an increasing function of W^o, the wealth elasticity of demand for money is *at least one*; this generalizes Arrow's result to the case of two unsafe assets and $\rho < 0$.

If, when $\rho < 0$, the RRA is greater than one, the contribution of the hedging effect to the demand for bonds is negative (and, hence, a positive contribution to the demand for money); the properties (a), (b) and (c), above, change sign whereas (d) and (e) remain of the same sign. Notice, however, that, under these circumstances, there is some ambiguity with respect to the overall effect on the demand for bonds as a result of a change in the riskiness of the nominal yield on bonds. A change in $\sigma_{\Pi B}$ causes the speculative demand to decrease but the

hedging demand increases:

$$\frac{\partial A^*}{\partial \sigma_{\Pi B}} = -\frac{1}{RRA}\cdot\frac{\bar{\Pi}_B 2\sigma_{\Pi B}}{(\sigma_{\Pi B}^2)^2} - \frac{\rho\sigma_{\Pi X}}{\sigma_{\Pi B}^2}\left(1-\frac{1}{RRA}\right) < 0 \text{ if}$$

$$\frac{1}{RRA}\frac{\bar{\Pi}_B 2\sigma_{\Pi B}}{(\sigma_{\Pi B}^2)^2} > -\frac{\rho\sigma_{\Pi X}}{\sigma_{\Pi B}^2}\left(1-\frac{1}{RRA}\right)$$

As $\sigma_{\Pi B}$ is, as reference to (8) and (11) will show, an argument in the expected opportunity of holding cash, the result indicates that an increase in that opportunity cost might lead to increased holdings of cash. Appendix B will explain this result.

Case C, $\rho > 0$: In this case, bonds are a positive hedge against depreciation in the value of the dollar. The properties of the hedging effect are symmetrical to those presented for $\rho < 0$; I look, however, very briefly at the case when the RRA is greater than one.

(a) $$\left.\frac{\partial A_H^*}{\partial \sigma_{\Pi X}}\right|_{1<RRA<\infty,\rho>0} > 0$$

(b) $$\left.\frac{\partial A_H^*}{\partial |\rho|}\right|_{1<RRA<\infty,\rho>0} > 0$$

(c) $$\left.\frac{\partial A_H^*}{\partial \sigma_{\Pi B}}\right|_{1<RRA<\infty,\rho>0} < 0$$

The above properties are intuitively plausible and need no further explanation. Additional properties are

(d) $$\left.\frac{\partial A_H^*}{\partial (RRA)}\right|_{1<RRA<\infty,\rho>0} > 0$$

Thus, if bonds are a positive hedge against inflation, the hedging demand for bonds increases as the household becomes more risk averse; however, the speculative demand decreases with an increase in the RRA. The overall effect on the demand for bonds is negative if $\bar{\Pi}_B > \sigma_{\Pi XB}$. This latter condition is readily shown to be necessary for portfolio diversification to be efficient.[16] The result thus indicates that at the margin the positive risk premium associated with holding bonds is not sufficient to compensate for an increased aversion to risk and that consequently the household substitutes out of bonds into money.

(e) $$\left.\frac{\partial A_H^*}{\partial W^o}\right|_{1<RRA<\infty,\rho>0} = \frac{\partial A_H^*}{\partial (RRA)}\frac{d(RRA)}{dW^o} \gtrless 0$$

where $d(RRA)/dW^o \gtrless 0$. As, in (d) above, the partial derivatives of the speculative and hedging demands are of opposite sign; again as in (d) above, if $d(RRA)/dW^o > 0$, the overall effect on the household's demand for bonds is negative if $\bar{\Pi}_B > \sigma_{\Pi XB}$. This result relies only on the RRA being positive. Thus, if diversification is efficient, the wealth elasticity of the demand is *at least one* if the RRA is an increasing function of W^o. Arrow's result is therefore general to the two unsafe asset case.

The hedging effect contributes negatively, when $\rho > 0$, to the speculative demand if the RRA is less than one; the properties (a), (b), and (c) above change sign, whereas (d) and (e) remain of the same sign. Again as for the case $\rho < 0$ and $RRA > 1$, there is an

[16]Because the analysis has used a mean-variance approximation to the expected utility hypothesis the need for such a condition is apparent. It is derived as follows: the expected value of the portfolio is, under the assumption that $W^o = P_X^o = 1$,

(a) $$\mu = E[W/P] = E[1 + r_M + A(r_B - r_M)]$$

where A is the percentage of the portfolio in bonds. The variance, as shown in Appendix B, is

(b) $$\sigma^2 = V[W/P] = A^2\sigma_{\Pi B}^2 + \sigma_{\Pi X}^2 - 2A\sigma_{\Pi XB}$$

Minimization of (ii) subject to $E[W/P] = \bar{\mu}$, a parametric value, yields the efficiently locus, the slope of which, $d\sigma^2/d\mu$, evaluated at $E[W/P] = \bar{\mu}$, is positive as required if

(c) $$\frac{\bar{\mu} - 1 - \sigma_{\Pi X}^2 + \bar{\Pi}_X}{(\bar{\Pi}_B - \sigma_{\Pi XB})} > \frac{\sigma_{\Pi XB}}{\sigma_{\Pi B}^2}$$

Substituting A^*, as derived in the text, for A in (a) and substituting the result for $\bar{\mu}$ in (c) yields that, for positive RRA,

$$\left.\frac{d\sigma^2}{d\mu}\right|_{A=A^*} > 0 \quad \text{if} \quad \bar{\Pi}_B - \sigma_{\Pi XB} > 0$$

ambiguity for the overall effect on the demand for bonds as a result of a change in $\sigma_{\Pi B}$; the overall effect is negative, as in the previous case, when

$$\frac{1}{RRA}\frac{\bar{\Pi}_B 2\sigma_{\Pi B}}{(\sigma_{\Pi B})^2} > -\frac{\rho\sigma_{\Pi X}}{\sigma^2_{\Pi B}}\left(1-\frac{1}{RRA}\right)$$

IV

This paper has derived asset demand functions for a household concerned with protecting the real value of its wealth. It has shown that if the rate of inflation is uncertain that the decision maker will generally consider the hedging properties of an asset when determining portfolio composition. In particular it has shown that the portfolio demand for money need not rely on the speculative motive alone and that, indeed, even if the speculative motive were to reduce to zero the hedging motive might yield a positive holding of money balances. The analysis does not, nor can it, explain the transactions demand for money. The relevance of the analysis lies in the fact that it highlights the hedging effect and shows it, within an essentially simple, but not simplistic, framework, to be of some importance.

Appendix A: Notation

W^o	nominal wealth at the end of the first period
M	nominal balances
B	bonds
P^o_B	period-one price of a bond
X	composite good
P^o_X	period-one price of the composite good
$U = u(X)$	the direct utility function
$\bar{P}_X = E[P_X]$	mean of the random variable P_X
$\bar{P}_B = E[P_B]$	mean of the random variable P_B
σ^2_X	variance of P_X
σ^2_B	variance of P_B
σ_{XB}	covariance between P_X and P_B
W	uncertain nominal value of the portfolio
$\bar{\Pi}_X = E[\Pi_X]$	mean of the random variable Π_X, the rate of inflation
$\sigma^2_{\Pi X} = V[\Pi_X]$	the variance of Π_X
$\bar{r}_M = E[r_M]$	mean of the random variable r_M, the real rate of return on a unit of nominal balances
$\bar{r}_B = E[r_B]$	mean of the random variable r_B, the real rate of return on a dollars worth of bonds
$\bar{\Pi}_B = E[\Pi_B]$	mean of the random variable Π_B, the nominal yield on bonds
$\sigma^2_{\Pi B} = V[\Pi_B]$	the variance of Π_B
$\sigma_{\Pi XB}$	the covariance between Π_X and Π_B
X^d	the demand for X
$H(W, P_X)$	the indirect utility function
H_i	the partial derivative of the H function with respect to the ith argument.
B^*	the portfolio demand for bonds
$RRA = \frac{-H_{11}W^o}{H_1}$	index of relative risk aversion
A^*	demand for bonds as a fraction of present wealth
M^*	the portfolio demand for money
N^*	demand for bonds under the assumption of certainty about the future rate of inflation
ρ	correlation coefficient between Π_X and Π_B
$\sigma_{\Pi X}$	standard deviation of Π_X
$\sigma_{\Pi B}$	standard deviation of Π_B
A^*_H	hedging demand for bonds
$(\bar{r}'_M - r'_B)$	net expected return to a

	unit of nominal balances when $\Pi_B = 0$
ϕ	the risk premium
σ^2	the variance of the portfolio
A	proportion of the portfolio held in bonds

APPENDIX B:[17] THE INTERRELATIONSHIP BETWEEN RETURNS, RISK, AND THE INDEX OF RISK AVERSION

It is useful, for the purpose of this Appendix, to rewrite the demand for bonds equation as

$$\text{(A1)}\quad A^* = \frac{\Pi_B}{\sigma^2_{\Pi B}} - \frac{(\Pi_B - \sigma_{\Pi XB})}{\sigma^2_{\Pi B}}\left(1 - \frac{1}{RRA}\right) = \frac{\Pi_B}{\sigma^2_{\Pi B}} - \frac{(\bar{r}_B - \bar{r}_M)}{\sigma^2_{\Pi B}}\left(1 - \frac{1}{RRA}\right)$$

where $\Pi_B/\sigma^2_{\Pi B}$ is the household's demand for bonds if the $RRA = 1$. Assume (i) $\rho < 0$ such that $(\bar{r}_B - \bar{r}_M)$ is, by (8) and (11) of the text, unambiguously positive and (ii) that the $RRA > 1$. The second term on the right-hand side of (A1) thus adds negatively to A^*. If, under these circumstances, ρ decreases such that $(\bar{r}_B - \bar{r}_M)$, the expected opportunity cost of holding money, increases the household's reaction is to demand less bonds. It is the purpose of this Appendix to clear up that apparent anomaly—and others like it, which are readily derived.

The household's demand for bonds follows, as differentiation readily shows, expected behavior with respect to changes in Π_B. The anomaly thus results from the covariance term and the associated hedging effect. The analysis, therefore, is pursued under the assumption that $\Pi_B = 0$. This yields

$$\text{(A2)}\quad A^* = -\frac{\sigma_{\Pi XB}}{\sigma^2_{\Pi B}}\left(1 - \frac{1}{RRA}\right) = \frac{(\bar{r}'_M - \bar{r}'_B)}{\sigma^2_{\Pi B}}\left(1 - \frac{1}{RRA}\right)$$

[17] I am indebted to Russell Davidson for noting and correcting errors in an earlier version of this Appendix: his intervention resulted in the present form of the Appendix.

where $(\bar{r}'_M - \bar{r}'_B)$ is the net expected return to a unit of nominal balances when $\Pi_B = 0$. Now, when $\rho > 0$ and the $RRA > 1$ an increase in $(\bar{r}'_M - \bar{r}'_B)$ will cause A^* to *increase*, a similar anomaly to the above, and the one that will be analyzed. All the other anomalies are variants of the above and are explained by the same rationale as given below.

Notice, first, that in the above circumstances the net real return to money, $(\bar{r}'_M - \bar{r}'_B)$ is positive and that money is a riskier asset than bonds. It will be shown that the anomaly arises, for an individual whose RRA is greater than one, because an increase in $(\bar{r}'_M - \bar{r}'_B)$ is accompanied by an unacceptably large increase in the riskiness of money. To make this precise I employ the Arrow-Pratt interpretation of the RRA. In this interpretation the RRA is measured in the neighborhood of some certain prospect. Since, within the content of the problem under discussion, the only certain commodity could be X (which, as previously noted, the individual is not permitted to hold) *one* unit of X is chosen as the certain prospect. Then the RRA is twice the risk premium ϕ, per unit of variance σ^2, (both ϕ and σ^2 measured in units of X) which must be offered with a risky portfolio and the certain prospect. Thus for such a risky portfolio.

$$\text{(A3)}\quad \phi = 1/2\, RRA\, \sigma^2$$

If, as is assumed, the RRA is locally constant then for a small change in the individual's portfolio he is left indifferent or is made better off according as

$$\text{(A4)}\quad d\phi = 1/2\, RRA\, d\sigma^2$$

or

$$\text{(A5)}\quad d\phi > 1/2\, RRA\, d\sigma^2$$

For simplicity of argument assume that

$$P^o_B = P^o_X = W^o = 1$$
$$\Pi_B = \Pi_X = 0$$

It is perhaps worth noting that had the individual been allowed to hold X then, from the above, 1 unit of X would be his certain prospect.

Let that proportion of the portfolio held in bonds be denoted by A. Then from equa-

tions (8) and (11), in the text, the expected real rate of return to the portfolio is

$$\text{(A6)} \qquad \phi = \sigma^2_{\Pi X} - A\rho\sigma_{\Pi X}\sigma_{\Pi B}$$

The variance is computed as follows

$$\begin{aligned}\text{(A7)} \quad \sigma^2 &= var\left(\frac{W}{P_X}\right) - var\left(\frac{1-A}{P_X} + \frac{AP_B}{P_X}\right)\\ &= var\left(\frac{1+A\Pi_B}{1+\Pi_X}\right)\\ &= var(1 + A\Pi_B - \Pi_X)\\ &\quad \text{(for sufficiently concentrated distributions of } \Pi_X \text{ and } \Pi_B)\\ &= A^2\sigma^2_{\Pi B} + \sigma^2_{\Pi X} - 2A\rho\sigma_{\Pi X}\sigma_{\Pi B}\end{aligned}$$

A small move into bonds, dA, will consequently produce the following change in ϕ and σ^2

$$\text{(A8)} \qquad d\phi = \frac{\partial\phi}{\partial A}dA = -\rho\sigma_{\Pi X}\sigma_{\Pi B}dA$$

$$\text{(A9)} \qquad d\sigma^2 = \frac{\partial\sigma^2}{\partial A}dA = (2A\,\sigma^2_{\Pi B} - 2\rho\sigma_{\Pi X}\sigma_{\Pi B})dA$$

It follows from (A7) that this move benefits the individual iff

$$\text{(A10)} \qquad \rho\sigma_{\Pi X}\sigma_{\Pi B} < 1/2\,RRA \cdot (2\rho\sigma_{\Pi X}\sigma_{\Pi B} - 2A\,\sigma^2_{\Pi B})$$

From (8) and (11), in the text, the left-hand side of (A10) is the net expected real return to one dollar; the bracketed expression on the right-hand side of (A10) is the increase in the variance from holding an extra dollar, that is, the marginal riskiness of moving out of bonds into money. There are three ways that $(\bar{r}'_M - \bar{r}'_B)$ can increase, viz., via ρ, $\sigma_{\Pi X}$, $\sigma_{\Pi B}$: the case of ρ is examined: the rest are analogous.

Imagine that the individual has selected his optimal portfolio for some value of ρ, $\sigma_{\Pi X}$, $\sigma_{\Pi B}$, i.e., his chosen proportion of bonds, A^*, equates the two sides of (A10). Now let ρ increase by an amount $d\rho$. Then the increase in $(\bar{r}'_M - \bar{r}'_B)$ is $\sigma_{\Pi X}\sigma_{\Pi B}d\rho$ and the increase in the marginal riskiness of moving is $2\sigma_{\Pi X}\sigma_{\Pi B}d\rho$. If the individual does *not* move from A^* inequality (A10) holds iff the RRA is greater than one. A move into bonds would thus benefit him: the risk premium foregone in the move is more than compensated by the decrease in risk.

REFERENCES

Kenneth J. Arrow, *Aspects of the Theory of Risk Bearing*, Helsinki 1964.

N. Biger, "The Assessment of Inflation and Portfolio Selection," *J. Finance*, May 1975, *30*, 451–67.

F. Black, "Capital Market Equilibrium with Restricted Borrowing," *J. Bus., Univ. Chicago*, July 1972, *45*, 444–55.

C. F. J. Boonekamp and D. Donaldson, "Biased Measurements with Unbiased Expectations," *Econ. J.*, Dec. 1977, *87*, 755–60.

W. E. Diewert, "Applications of Duality Theory," in Michael Intrilligator and David Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam 1974.

B. Eden, "On the Specification of the Demand for Money: The Real Rate of Return versus the Rate of Inflation," mimeo, Univ. Chicago, Oct. 1974.

S. Fischer, "The Demand for Index Bonds," *J. Polit. Econ.*, June 1975, *83*, 509–34.

G. Hanoch, "Desirability of Price Stabilization," disc. paper no. 351, Harvard Instit. Econ. Res., Mar. 1975.

R. A. Jones, "Price Uncertainty and the Use of Money as a Standard of Deferred Payment," unpublished paper, Univ. California-Los Angeles, Mar. 1975.

H. Markowitz, "Portfolio Selection," *J. Finance*, Mar. 1952, *7*, 77–91.

R. Roll, "Assets, Money and Commodity Price Inflation under Uncertainty," *J. Money, Credit, Banking*, Nov. 1973, *5*, 903–23.

P. A. Samuelson, "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments," *Rev. Econ. Stud.*, Oct. 1970, *37*, 537–42.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.

———, "The Theory of Portfolio Selection," in Frank H. Hahn and Frank P. Brechling, eds., *The Theory of Interest Rates*, New York 1966.

The Effect of Unemployment Insurance on Temporary Layoff Unemployment

By MARTIN FELDSTEIN*

Economists are now beginning to recognize that an understanding of temporary layoffs is crucial for a proper analysis of unemployment. In manufacturing, about 75 percent of those who are laid off return to their original employers. More generally, among all persons classified as "unemployed job losers," temporary layoffs account for about 50 percent of all unemployment spells. Temporary layoffs are an even larger fraction of cyclical changes in the number of job losers. While this group includes some seasonally unemployed, most temporary layoffs are induced by short random or cyclical fluctuations in demand. The conventional model of search unemployment is inappropriate for those on temporary layoff and the modern theory of the Phillips curve requires substantial modification because of the size and cyclical variation of temporary layoff unemployment.[1]

In a previous paper (1976), I showed analytically that our current system of unemployment insurance (*UI*) provides a substantial incentive for increased temporary layoff unemployment.[2] The present paper provides micro-economic evidence that *UI* actually has such a powerful effect. The estimates imply that the incentive provided by the current average level of *UI* benefits is responsible for approximately one-half of temporary layoff unemployment.

It is important to note that the current study shows that *UI* increases the *amount* of temporary layoff unemployment, but does not deal with the *mean duration* per spell. This distinction deserves emphasis because nearly all previous empirical work focused on the potential effect of *UI* on duration. This focus on duration is both unfortunate and surprising since *UI* can actually increase total unemployment while decreasing the mean duration per spell. While *UI* increases the duration of any *given* spell of unemployment, it may also induce more very short spells of unemployment. This possibility of reduced mean duration is clear in my 1976 theoretical analysis. An additional practical

*Professor of economics, Harvard University. I am grateful to the National Science Foundation for support of this research, to David Ellwood and Joseph Kahan for assistance with the statistical calculations, and to Richard Freeman, Zvi Griliches, Daniel Hamermesh, James Medoff, Melvin Reder, and Jeffrey Sachs for discussions and comments. Earlier versions of this paper were presented at seminars at Chicago, Harvard, and Yale universities.

[1] In my 1975 paper, pp. 737–42, I discuss the implications of temporary layoffs for the theory of search unemployment, the Phillips curve, and wage inflexibility. Although the standard criterion of unemployment is active job seeking within the past four weeks, individuals are officially classified as unemployed without any inquiry about recent job-seeking activity if they state that they are "on layoff awaiting recall by their employers." Some of those on layoff look for temporary jobs or alternative permanent employment, but the vast majority do return to their original employers. Readers should not be confused by the two quite separate meanings of the term "layoff" in the Department of Labor's lexicon. In manufacturing establishment data, a layoff is a separation initiated by the employer (not a quit) and may be permanent or temporary. In the *Current Population Survey* (*CPS*), an individual is on layoff if he is not working but "has a job" to which he is expecting to be recalled by his employer. To emphasize that I am dealing with those layoffs expected to terminate in recall, I use the adjective "temporary." Unfortunately, the *CPS* uses the word temporary in a different and quite confusing way: persons on layoff are divided into an "indefinite duration" group (in which the individual does not have an expected date of recall within thirty days) and a "temporary" group (when such a date is known). When it is useful to distinguish these groups, I use the terms "indefinite duration" and "fixed duration"; in my usage, the term temporary layoff includes both groups.

[2] My 1976 paper is really an explicit proof of arguments made more informally in my earlier study for the Joint Economic Committee (1973). For a similar development, see Martin Baily.

matter reinforces this tendency. In the absence of *UI*, firms might be reluctant to lay off workers for short periods in response to random demand fluctuations, for fear of losing these workers to other firms, or at least of creating costly ill will; *UI* eliminates these problems and facilitates short-duration layoffs. In contrast, firms might have no choice but to lay off employees for long spells during the less frequent, but more protracted, spells of low demand. Unemployment insurance thus increases the number of spells of temporary layoff unemployment with a relatively greater increase for short spells. Since the duration should increase for any given spell, while the mix should change to add short spells that would otherwise not exist, the net effect of *UI* on duration is indeterminate. The existing estimates of the effect of *UI* on the mean duration of unemployment spells should therefore be regarded as an understatement—and, possibly, an extreme understatement—of the effect of *UI* on total unemployment.[3]

Although the presence of a labor union is not necessary to obtain the effects on temporary layoffs indicated by the theoretical analysis of *UI*, these predicted effects are likely to be magnified if the employees are unionized.[4] The basic reason is that employers are more willing to lay off workers when they are confident that they will return when recalled, while employees are more willing to be laid off if they can be confident that they will be recalled. Both conditions are more likely to be met in unionized firms where workers often receive compensation that exceeds their market alternative, and have seniority privileges and pensions that are not portable. More directly, union contracts often guarantee that previous workers will be recalled before any new employees are hired (see U.S. Bureau of Labor Statistics, 1972). Finally, unionized firms may have more layoff unemployment because, as Freeman has suggested, unions provide an effective mechanism for expressing workers' collective preferences to management. All of this implies that temporary layoff unemployment should be higher for union members and suggests that the response of temporary layoff unemployment to *UI* benefits may also be greater.

The first section of this paper discusses the data and methods used in the present study. The econometric estimates are presented in Section II. The brief concluding section suggests some directions for future analysis and comments on the implications of the research for the optimal redesign of social insurance.

I. Data and Method

The current study uses a sample of nearly 25,000 individual observations collected by the *Current Population Survey* (*CPS*) to measure the effect of unemployment insurance benefits on temporary layoff unemployment. The estimated regression equations presented in the next section relate each individual's temporary layoff unemployment status (a binary variable equal to 1 if the individual is on temporary layoff) to three kinds of variables: 1) his potential *UI* benefit as a percentage of lost net wages; 2) his basic demographic characteristics; and 3) the basic characteristics of his employment. This section begins by discussing the *CPS* sample and the method of calculating the potential *UI* benefit "replacement ratio" for each individual. The measurement of demographic and employment characteristics is then discussed.

A. *The CPS Sample*

The *CPS* is the government household survey used by the Department of Labor to calculate official monthly unemployment

[3]This criticism applies more generally to those (like Stephen Marston) who measure the effect of *UI* on the average duration of all types of unemployment. Since *UI* is expected to induce additional temporary layoffs and the mean duration of temporary layoff unemployment spells is substantially less than that of other types of unemployment (see the author, 1975), *UI* may actually reduce the overall mean duration while increasing both total unemployment and the duration of every spell that would have existed without unemployment insurance.

[4]This paragraph reflects discussion with my colleagues Richard Freeman and James Medoff; see Medoff.

rates. About 60,000 households are interviewed each month about the employment activities of their members during the week prior to the survey. The March survey of each year also obtains information about labor force participation, employment, and earnings during the previous year. The current study uses the survey for March 1971, a period of relatively high unemployment.[5]

For this analysis, individuals were eliminated from this *CPS* sample if they were not in the experienced labor force, were reentrants to the labor force, or were self-employed; none of these groups is at risk of being laid off. Also eliminated because of the atypical character of their employment were employees in the public sector and in agriculture.[6] To avoid the problems associated with those who combine school and work, and with those on the verge of retirement, the sample was restricted to individuals between the ages of 25 and 55. Finally, a few observations were excluded, because data were missing on the individual's age, sex, color, marital status, industry and occupation of employment, union membership, or previous year's work experience. The sampling weights indicate that the resulting sample of 24,545 represents a population of 34.2 million persons.[7]

B. *Calculation of Potential UI Benefits*

The unemployment compensation benefits for which an individual is eligible depend on his previous earnings up to a ceiling of maximum benefits received by about half of all *UI* benefit recipients. Because unemployment insurance is actually a series of state programs that operate as part of a general federal system, the formulae relating benefits to past earnings and the maximum benefits differ among the states. In addition, dependents' benefits are also available in states with approximately one-third of covered workers.

The *CPS* collects no information about the unemployment insurance benefits received by the currently unemployed or the potential benefits of the employed.[8] A special computer program was therefore prepared to evaluate the potential *UI* benefits for each of the 24,545 individuals in the final *CPS* sample. The algorithm uses the particular rules for each individual's state of residence and incorporates information on his industry of employment, previous year's earnings and work experience, and number of eligible dependents.[9] As a rough test of the accuracy of this method, the program was used to determine the benefit eligibility and to calculate the benefits for *all* unemployed persons in the full *CPS* sample (and not the final subsample of 24,545 observations). The implied total benefits for March 1971 was $540 million; this is reasonably close to the total amount actually paid as reported by the individual state *UI* agencies, $630 million. The accuracy is likely to be greater for temporary layoffs for whom the reporting of previous year's income is much more reliable.

[5]The seasonally adjusted unemployment rate in March 1971 was 6.0 percent and had been stable during the previous three months. The March 1971 survey was not "selected," but was the first *CPS* tape that became publicly available. The use of that sample indicates the slow gestation of this project.

[6]Barry Chiswick presents evidence that the recent extension of unemployment insurance to agriculture has substantially increased the seasonality of employment and unemployment in agriculture. It will be important to see if that result is confirmed by data after the 1975 recession year.

[7]In 1971 there were 50.8 million persons in the labor force between the ages of 25 and 55. The difference between 50.8 million and 34.2 million represents primarily government employees, agricultural workers, and the self-employed.

[8]Individuals are asked about the total annual value of benefits received during the previous year, but these twelve-month recall data are notoriously bad and, in the aggregate, represent a 50 percent understatement of the amounts paid by the *UI* program.

[9]There is no information on "benefit exhaustion," i.e., on whether an individual has already been unemployed so long that his number of weeks of eligibility for benefits has been exhausted. However, for all types of unemployment, only about 20 percent of spells exhaust available benefits while, for those on temporary layoff, the percentage should be very much lower: in March 1974, only 4 percent of "job losers on layoff" were unemployed for more than twenty-six weeks, while 12 percent of "job losers with no job" were unemployed for that long (see the author, 1975, Table 4).

The central variable of interest is the ratio of the individual's potential unemployment insurance benefit to his foregone earnings net of marginal income and payroll taxes. This *UI* "benefit replacement ratio" measures the proportion of lost net-of-tax earnings that would be replaced by *UI* benefits. A 60 percent *UI* benefit replacement ratio implies that the unemployed individual would lose only 40 percent of his previous net-of-tax wage income. Stated differently, the benefit replacement ratio is analogous to a rate of tax levied on earnings when the alternative is insured unemployment; a 60 percent benefit replacement ratio implies that the individual, by working instead of collecting *UI*, receives additional income equal to only 40 percent of his total net wage. The computer program evaluated the benefit replacement ratio for each individual, using the federal income tax schedules, to evaluate a marginal tax rate for someone with the individual's family income and dependents who used the standard deduction. The relevant marginal social security tax rate and state income tax rate were added to the federal marginal tax rate.

Although theory predicts that the probability of being on temporary layoff is an increasing function of the benefit replacement ratio, there is no presumption of linearity. A movement in the benefit replacement ratio from 0.70 to 0.80 may increase unemployment by more than a movement from 0.30 to 0.40. To eliminate the restriction of a linear specification, equations are reported in the next section in which the continuous benefit replacement ratio variable (BEN) is replaced by a set of binary variables that classify individuals by their benefit replacement ratios: $BEN = 0$ (for those not eligible for benefits); $0 < BEN \leq 0.30$; $0.30 < BEN \leq 0.50$; $0.50 < BEN \leq 0.70$; $0.70 < BEN \leq 0.85$; and $0.85 < BEN$. This method has the further advantage that it can clearly separate those who are ineligible for benefits ($BEN = 0$) from the remaining variation in BEN.

Although I believe that this represents the best method of evaluating the benefit replacement ratio with the available data, there are several problems that should be borne in mind in evaluating the results. First, there is no information on the extent of experience rating that is relevant for each individual's employer. If the extent of experience rating is uncorrelated with the benefit replacement ratio, ignoring experience rating does not bias inferences about the effect of the benefit replacement ratio on temporary layoff unemployment.[10] Second, there are three omissions that are likely to cause an overestimate of the impact of unemployment insurance on temporary layoff unemployment: cash and in-kind transfers that may be available to individuals on temporary layoff, the value of fringe benefits that are lost during unemployment, and the work expenses (transportation, meals, etc.) that are avoided during unemployment. None of these omissions is likely to be large for the quite short duration of unemployment that are relevant here. Moreover, to the extent that a higher probability of layoff is compensated by a higher gross wage (as implied by the firm's budget constraint), there will be an offsetting underestimate of the impact of *UI* on temporary layoff unemployment. It is difficult to assess the net effect of these countervailing influences, but the resulting bias is likely to be small.

It is much more important to understand that the regression coefficient of the benefit replacement ratio measures the effect of *interindividual differences* in unemployment benefits and that the effect of such differences is less than the effect of a *general* change in everyone's benefit replacement

[10]The regression of the unemployment variable on the benefit replacement ratio does, however, understate the effect on unemployment of differences in the net *UI subsidy*. The net *UI* subsidy is the difference between the benefits and the additional experience-rated tax payments induced by those benefits. In the notation of my 1976 paper, the net subsidy is $[1 - e(1 - t_y)]b$ where e is the ratio of induced employer tax to incremental benefits (i.e., the extent of experience rating), by the marginal personal income tax rate, and b is the weekly benefit. If $e(1 - t_y)$ were constant, the regression coefficient of b would understate the effect of changes in the net subsidy by a factor of $[1 - e(1 - t_y)]^{-1}$.

ratio. As a general rule, it is the employer who makes the decision to lay off and recall a worker, while the employee himself is essentially passive.[11] An employer can respond to his employees' benefit replacement ratios only as an average for the group whose layoff he is considering and not individually for each member in the group. It is because the relevant group of employees within a firm has similar benefit replacement ratios[12] that the individual benefit replacement information is relevant for understanding what is essentially an employer or employer-employee group decision. Since the benefit replacement ratios are not identical for the relevant group of the firm's employees, some part of the variation of *BEN* in the sample will not affect layoff unemployment. The effect of this is to make the estimated regression coefficient an underestimate of the effect of a *general* increase or decrease in all benefits.[13]

C. *Demographic and Employment Characteristics*

The demographic characteristics included in the analysis are the standard list of age, sex, marital status, and race.[14] As I indicated above, separate equations are also estimated for men only. The sample is limited to individuals between the ages of 25 and 55. To avoid any assumption about the form of the relation between age and temporary layoffs, individuals are divided into four separate age groups and binary variables are used in the regression equation. The age groups included are 25–29, 30–39, and 40–49; the coefficient for persons 50–55 is implicitly zero. The other demographic variables are self-explanatory.

The potential role of unions was discussed briefly in the introduction. In the final *CPS* sample of 24,545, 6,845 individuals (or 27 percent of the sample) indicated that they were members of labor unions. There is no indication whether the individual's current employment is in a union job. This suggests that the estimated coefficient of the union variable may underestimate the full effect of unionization.

Individuals were classified according to industry group and occupation category and the corresponding binary variables were included in the regression to control for inherent "technological" differences among them in the likelihood of layoffs.[15] This procedure entails a danger of "overcontrolling" for the exogenous aspect of these variables. Individuals with high potential benefit replacement rates (for example, with low wage rates or high spouse income, or large families in states where dependents' allowances are paid) may seek employment in industries and occupations with high technological probabilities of layoff unemployment. To the extent that this is important, the regression coefficients will overstate the importance of the industry and occupation variables and will understate the impact of the benefit replacement ratio. Although it is not possible to model

[11] I say "as a general rule" because workers do frequently have "inverse seniority" privileges that permit more senior workers to *choose* to be laid off before or instead of others. See U.S. Bureau of Labor Statistics (1972) for a description of these privileges.

[12] The benefit replacement ratios are similar to the extent that members of the group have similar wages and, being located in the same state, have similar unemployment benefit schedules and state tax rates.

[13] This can be stated differently by noting that a firm can only perceive and respond to the mean *BEN* value for the relevant group of its employees and essentially ignores the *within-group variance.* A general change in all *UI* benefits shifts this mean while part of the sample variation includes the within-group variance. In still different language, the coefficient of *BEN* is biased down but the size of the bias is limited to the extent that the between-group variance is large relative to the within-group variance. This bias can be thought of as a classical "errors-in-variables" bias: the "true" value of *BEN required by the model* is *the mean* of the individual *BEN* values for the relevant employee group, while the *actual* individual *BEN* values may be regarded as equal to the "true" value plus an error. This errors-in-variables interpretation also indicates that there is a downward bias that is an increasing function of the within-group variance relative to the between-groups variance.

[14] It might be interesting to extend this list to other attributes that reflect differences in tastes for leisure, for example, education, home ownership, age of children of married women, etc. See, however, fn. 13.

[15] The twelve industry groups were combinations of two-digit industries. Recall that agricultural workers, the self-employed, and public employees were omitted from the sample. The nine occupation groups were combinations of more detailed two-digit classifications.

this simultaneous relationship, separate results will be presented with and without the industry and occupation variables.

The final variable considered in the analysis is the individual's wage rate. Temporary layoff unemployment is likely to be related to the individual wage in several quite different ways. First, for any given benefit replacement ratio, a higher wage implies both a higher absolute benefit and a greater absolute cost of unemployment; the sign of this effect is therefore indeterminate. Second, if a high wage reflects better pay relative to the individual's market opportunity,[16] the employer will be more likely to lay off workers with a confidence that they will return when recalled; this implies a positive coefficient for the wage variable. Third, a higher wage may *ceteris paribus* imply greater seniority; greater seniority means fewer involuntary layoffs relative to other employees within the firm, but a group with more seniority on average may have more temporary layoffs because workers are more likely to await recall.

Related to this seniority aspect is the possibility that more senior workers who are laid off perceive themselves (correctly) as only on temporary layoff, while their more junior coworkers who are laid off may regard the separation as permanent because their probability of recall is substantially lower. Finally, jobs with more layoffs may pay higher wage rates *ceteris paribus* than other jobs, implying that the gross wage is endogenous and positively related to the unemployment probability. While this source of wage variation is likely to be small relative to the wage variation reflecting individual skill differences, etc., some equations without this variable have been estimated to assess the effect of erring in the direction of its omission.

II. The Econometric Evidence

All of the equations that I have estimated imply that the current level of unemployment insurance benefits causes a substantial fraction of the observed temporary layoff unemployment. More specifically, the econometric evidence indicates that the temporary layoff unemployment rates would be reduced by approximately one-half if the adverse incentive provided by the current unemployment insurance were eliminated. This conclusion is not sensitive to the exclusion of questionable regressions or to the restriction of the sample to particular subsamples.[17]

Before looking at the estimated regression coefficients, it is helpful to examine the basic data on temporary layoff unemployment rates and *UI* benefit replacement ratios. In March 1971 the temporary layoff unemployment rate was 1.6 percent in the population corresponding to the final *CPS* sample of 24,545 employees; that is, on average, the corresponding population had a probability of 0.016 of being unemployed and on layoff during the sample week in 1971. The mean value of the benefit replacement ratio for this population was 0.55. Only 3 percent of the population was found to be ineligible for benefits[18] while 60 percent of the sample had benefit replacement rates above one-half and 30 percent had benefit replacement rates about 70 percent.[19]

Table 1 shows the temporary layoff unemployment rates corresponding to six levels of the benefit replacement ratio. This unemployment rate rises monotonically from 0.50 percent among the ineligibles ($BEN = 0$) to 2.17 percent in the highest benefit group ($BEN > 0.85$). Taken at face value, these unemployment rates imply that

[16] Recall that we are "holding constant" the effect of age, sex, color, unionization, industry, and occupation. It might be interesting to add education and other variables.

[17] The reader should remember the caveats and potential biases discussed in Section I; they will not be repeated here.

[18] Recall that the sample is restricted to eliminate many groups with no *UI* benefits, such as new entrants and the self-employed.

[19] The distribution of benefit replacement ratios for the population should not be confused with the distribution for the unemployed subgroup. The mean benefit replacement ratio of the unemployed was 0.59; if those with zero benefits are excluded, the mean benefit replacement rate for the eligible unemployed exceeds 0.60. This is consistent with the calculation that I presented for a range of hypothetical employees in my 1974 paper.

TABLE 1—UNEMPLOYMENT INSURANCE AND TEMPORARY LAYOFF UNEMPLOYMENT[a]

UI Benefit Replacement Ratio (*BEN*) (1)	Percentage of Population (2)	Temporary Layoff Unemployment Rate (3)
0	3.3	0.50 (0.25)
$0 < BEN \leq 0.30$	8.4	1.26 (0.25)
$0.30 < BEN \leq 0.50$	27 9	1.30 (0.14)
$0.50 < BEN \leq 0.70$	30.0	1.80 (0.15)
$0.70 < BEN \leq 0.85$	23.4	1.83 (0.18)
$BEN > 0.85$	7.0	2.17 (0.35)
All Persons	100.0	1 60 (0.08)

[a]Based on the March 1971 *Current Population Survey* of 24,545 individuals. See text for definition of *BEN*. The figures in colums (2) and (3) are estimates of population rates based on *CPS* sampling weights. Approximate standard errors of the temporary layoff unemployment rates are shown in parentheses in column (3). (Note that these estimates are independent tabulations and not regression coefficients.)

reducing *BEN* to a maximum of 0.40 would lower the temporary layoff unemployment rate from 1.60 to 1.26, a reduction of 0.34 percentage points. It must, however, be borne in mind that this relation between benefits and temporary layoff unemployment rates is not adjusted for demographic or economic characteristics. We turn therefore to the multiple regression equations.

Table 2 presents the estimated coefficients of four basic regression equations. The dependent variable is binary, taking the value of 1 if the individual is unemployed and on layoff, and the value of 0 otherwise. The regression coefficients have all been multiplied by 100, converting the predicted dependent variable from a probability to a percentage unemployment rate. The sample means and proportions of the explanatory variables are shown in the first column.

Consider first the estimated coefficient of the benefit replacement ratio in equation (1).[20] The coefficient of 1.345 implies that the mean *BEN* value of 0.55 raises the mean temporary layoff unemployment rate by 1.345 x (0.55) = 0.74. Since the temporary layoff unemployment rate is 1.60, this equation implies that *BEN* is responsible for 46 percent of the observed temporary layoff unemployment rate. Because the industry and occupation variables may overcorrect for the truly exogenous effects of these variables, the basic specification is repeated without them as equation (2). The coefficient of the benefit replacement ratio rises slightly (to 1.545), implying that the mean benefit replacement ratio of 0.55 is responsible for 53 percent of the observed temporary layoff unemployment rate.[21]

The coefficient of the binary union variable in equations (1) and (2) provides strong evidence that union members are much more likely to experience temporary

[20]Note that equation (1) includes all of the variables discussed in Section I; the coefficients of the twenty-seven industry and occupation variables are not shown since they are not of interest in themselves, and would require much extra space in the table.

[21]Omitting the other potentially endogenous economic characteristic variables (the gross wage rate and unionization) only lowers this coefficient to 1.515. (This equation is not shown in the table.) Other variants cluster around 1.3, rising as high as 1.7 and falling as low as 1.0. Replacing the gross wage rate by a set of six classification variables in gross wages has essentially no effect on the other coefficients.

TABLE 2—DETERMINANTS OF TEMPORARY LAYOFF UNEMPLOYMENT

Variable	Sample Means and Proportions	Regression Coefficients (1)	(2)	(3)	(4)
BEN	0.55	1.345 (0.426)	1.545 (0.420)		
BEN = 0	0 03			−1.230 (0.539)	−1.484 (0.534)
0 < *BEN* ≤ 0 30	0 08			−1.552 (0.518)	−0.399 (0.516)
0 30 < *BEN* ≤ 0.50	0.28			−1.531 (0 391)	−1.343 (0.389)
0 50 < *BEN* ≤ 0.70	0.30			−1.074 (0 355)	−0.812 (0.356)
0 70 < *BEN* ≤ 0.85	0.24			−0.657 (0.344)	−0.490 (0.346)
Union	0.28	1.154 (0 204)	2.236 (0.183)	1.169 (0.204)	2.249 (0.183)
Age: 25-29	0.18	0 686 (0 270)	0.675 (0.270)	0.699 (0.270)	0.680 (0.270)
Age 30–39	0.31	0.220 (0.238)	0.235 (0.239)	0.229 (0.238)	0.241 (0.239)
Age 40–49	0 33	−0.196 (0.234)	−0.172 (0.235)	0.194 (0.234)	−0.168 (0.235)
Male	0.65	−1.460 (0.226)	−0.309 (0.195)	−1.279 (0.238)	0.189 (0.214)
Married	0.91	−0 267 (0.289)	−0.062 (0.289)	−0.243 (0 290)	−0.364 (0.291)
White	0.89	−0.068 (0 269)	−0.332 (0.261)	−0.068 (0.270)	−0.311 (0.262)
Gross Wage ($100)	1.64	0.202 (0 080)	0.127 (0 077)	.228 (0 094)	0.141 (0.092)
Industry-Occupation	-	a		a	-
Constant	-	2.025	0 333	3.687	1.931
Mean of Dependent Variable		1.601	1 601	1.601	1 601
N		24,545	24,545	24,545	24,545

Notes. All coefficients have been multiplied by 100, converting the dependent variable from a probability to a percentage unemployment rate Standard errors are shown in parentheses See text for description of sample and definitions of variables
[a]Indicates that the twenty-seven industry and occupation variables were included in the equation.

layoff unemployment than nonunion members. The temporary layoff unemployment rate is 1.15 percentage points higher than the rate for nonmembers even after adjusting for this industry-occupation mix. Without that adjustment, the differential is 2.24 percentage points. I will return below to the evidence that the layoff unemployment rate of union members is also more sensitive to *UI* benefits.

The coefficients of the other variables are interesting but involve no important economic insights. There is clear evidence that the frequency of temporary layoff unemployment falls quite sharply with age, a reflection of the powerful seniority system. There is no statistically significant difference between either whites and nonwhites or marrieds and singles. Males appear to have a significantly lower temporary layoff unemployment rate when (but only when) the industry and occupation effects are included separately.[22]

[22]The sex differential is large and surprising to me. It may be an artifact of overadjustment for industry and occupation or it may reflect a real difference between the sexes. Women may be more likely to take seasonal work (within broad industry-occupation groups) or to have relatively long spells of temporary layoff. Nothing is known about these fascinating issues

Equations (3) and (4) replace the continuous benefit replacement ratio variable by a set of six binary classification variables. In each equation, an increase in the benefit replacement ratio always implies an increase in the predicted temporary layoff unemployment rates.[23] Both equations suggest that variations in the benefit replacement ratio below the 30 to 50 percent range have little effect on unemployment but higher benefit replacement ratios have a substantial adverse effect. The coefficients of the *BEN* variables in equation (3) imply that lowering *BEN* for everyone to 0.40 (with an implicit coefficient of −1.53) would reduce the temporary layoff unemployment rate to 0.46 percentage points.[24] With equation (4), the same calculation implies a reduction of the temporary layoff unemployment rate of 0.49 percentage points (with an implicit baseline coefficient of −1.40). It is not clear how much weight should be given to the implications of this more elaborate specification. On purely statistical grounds, there is little basis for choice; the reduction in the residual sum of squares in going from equation (1) to equation (3) is not quite significant at the 5 percent level, while going from equation (2) to equation (4) is not even significant at the 10 percent level.[25] The pattern of the coefficients does correspond to the a priori ex pectation that variations in benefit replace ment ratios will have a weaker effect whe *UI* benefits are "too small to bother takin into account" than when those benefits re place a substantial fraction of lost net wag income. However, the apparently weak ef fect at low benefit levels may reflect onl the small fraction of the sample in thi range; since only 11 percent of the sampl had *BEN* values below 0.30, it is difficul to make any inferences about the effects o variations in benefits within the range be low 0.30 or between this range and the ne higher interval. It is probably best to re main agnostic on this question until mor data become available.[26]

Table 3 confirms that union membe have a substantially higher temporary la off unemployment rate and are more sens tive to unemployment insurance benefi than are nonunionized workers. For th sample of 6,845 union members, the ten porary layoff unemployment rate was 3.1 percent, twice the rate for the entire sampl and thus three times the rate for nonunio members.[27] The coefficient of the benef replacement ratio variable in equation (1) i 2.72, also about twice the correspondin coefficient for the entire sample.[28] A coeffi cient of 2.72 implies that the mean benef replacement ratio of 0.54 (for union mem bers) induces a 1.47 percent temporary la

and an adequate analysis would go beyond the proper focus of this paper.

[23]In equation (3) the step from "ineligible for benefits" to "eligible" appears to reduce temporary layoff unemployment. This implication should be given very little weight since the ineligible group is very small in the sample and the individuals who fall into that category are likely to have other special but unrecorded characteristics

[24]There would be an additional long-run reduction in the temporary layoff unemployment rate as production and employment shift out of the industries and occupations with high rates of temporary layoff that are currently subsidized by unemployment insurance.

[25]This *F*-test is only appropriate as an approximation since the continuous *BEN* variable is only approximately a linear combination of the binary *BEN* variables. Henri Theil's $\bar{R}^2$ criterion also indicates only the slightest possible preference for the more complex specifications. All of the R^2 values are extremely low, approximately 0.02; this is common for household survey data with a low probability binary dependent variable.

[26]The coefficients of the *BEN* variables in equatio (3) and (4) correspond quite closely to the condition unemployment rates presented in Table 1, for exampl an increase in *BEN* from 0.40 to 0.60 reduces the pr dicted temporary layoff unemployment rate by 0.5 percentage points in both the multiple regression equ tion and the unadjusted values of Table 1. It is cle from this comparison that replacing equations (3) ar (4) by logit regression instead of ordinary least squar would be very unlikely to change any of the concl sions of the current analysis.

[27]The total rate of 1.601 is a weighted average the union rate of 3.141 for 27.9 percent of the samp and 1.005 for the remaining 72.1 percent of nonunio members.

[28]The specification of equation (1) in Table 2 exactly the same as equation (1) in Table 1 exce that the union variable is now omitted. The coeff cients of the other variables are not shown in order save space.

TABLE 3—COEFFICIENTS OF BENEFIT REPLACEMENT RATIO VARIABLES FOR UNION MEMBER SUBSAMPLE AND MALE SUBSAMPLE[a]

Variable	Subsample Means and Proportions	Union Members Only (1)	(2)	(3)	(4)	Subsample Means and Proportions	Men Only (5)	(6)	(7)	(8)
BEN	0.54	2.723 (1.499)	2.287 (1.495)				1.419 (0.708)	1.584 (0.696)		
BEN = 0	0.01			−2.460 (2 468)	−2.968 (2.471)	0.01			−2.349 (1.316)	−2.544 (1.319)
0 < *BEN* ≤ 0.30	0 07			−4 382 (1.518)	−3.755 (1.516)	0.12			−1.564 (0.869)	−1.491 (0.867)
0.30 < *BEN* ≤ 0.50	0.36			−3.876 (1 193)	−3297 (1.192)	0.40			−1.470 (0.795)	−1.367 (0.794)
0.50 < *BEN* ≤ 0 70	0.33			−2.834 (1.126)	−2.394 (1.128)	0.33			−0.985 (0.787)	−0.811 (0 788)
0 70 < *BEN* ≤ 0.85	0 19			−2 508 (1.106)	−2.296 (1.110)	0.12			−0.759 (0.816)	−0.680 (0.818)
Includes Industry-Occupation Variables?	-	Yes	No	Yes	No		Yes	No	Yes	No
Mean of Dependent Variable	3 141	3.141	3.141	3.141	3 141		1.600	1.600	1 600	1.600
Sample Size	6,845	6,845	6,845	6,845	6,845		15,873	15,873	15,873	15,873

[a]Each equation also contains age, sex, color, marital status, and wage variables and a constant term (as in Table 1); their coefficients are not shown to save space. The Men Only equations also contain a union variable. The table indicates when industry and occupation variables are included. The omitted *BEN* category corresponds to *BEN* > 0.85 and has an implicit coefficient of zero.

off unemployment rate, or 47 percent of the overall 3.14 percent rate.[29]

Although the mean benefit replacement ratio for union members is almost exactly the same as for the entire sample, the distributions of benefit replacement ratios differ noticeably. The replacement ratios for union members are clustered more closely around the average; 69 percent of union members have *BEN* values between 0.30 and 0.70, while 58 percent of the entire sample is in this range. Almost no union members appear to be ineligible for benefits. The coefficients of equations (3) and (4) also show that the temporary layoff unemployment rate varies inversely with the benefit replacement ratio. Both equations imply that increasing the benefit replacement ratio from 0.40 to 0.60 raises the temporary layoff unemployment rate by about an entire percentage point.[30]

The results for the "men only" sample (presented in columns 5–8 of Table 3) are very similar to the estimates for the entire sample and need no detailed comment. The temporary layoff unemployment rate of 1.600 is almost identical to the rate for the entire sample (1.601). The regression coefficients differ substantially from the corresponding numbers of Table 1 only for the *BEN* = 0 subcategory; since only 0.01 percent of the men and 0.03 percent of the entire population are in this group, the comparison of the regression coefficients is without real substance.

Equations similar to those of Table 1 were also estimated with the "duration of

[29]Excluding the twenty-seven industry and occupation variables (as in equation (2)) reduces the coefficient slightly but leaves these conclusions essentially unchanged. The industry and occupation variables are themselves statistically significant so that equation (1) would be the clearly preferable specification except for the possible simultaneity problem noted in the text.

[30]The 1 percent of union members who are ineligible for benefits (*BEN* = 0) appear to have an unusually high layoff rate. This anomalous behavior also contributes to the relatively high standard error of the *BEN* variable in equations (1) and (2). The very small sample with *BEN* = 0 and correspondingly high standard error imply that no weight should be given to this group. For *BEN* > 0, equations (3) and (4) show a strong monotonic relation.

unemployment to date of survey" as the dependent variable. There was no significant relation between *BEN* and duration, implying that the effect of *UI* in inducing more short-duration spells of unemployment offset the effect of *UI* in lengthening the duration of spells that would otherwise have occurred.

Although I am tempted to compare the estimates presented in this section with the results of other recent studies, I believe that research is too dissimilar to warrant such comparison. There have been no previous econometric studies of the effect of unemployment insurance on temporary layoff unemployment. The recent econometric research has focused on the duration of unemployment spells or on the total unemployment rates for state aggregates.[31] There are several fascinating problems in the interpretation of these data, but their discussion belongs elsewhere.

III. Conclusion

The evidence presented in this paper implies that unemployment insurance has a powerful effect on temporary layoff unemployment. The average *UI* benefit replacement ratio implied by the current law can account for about half of temporary layoff unemployment. An increase in the *UI* benefit replacement ratio from 0.4 to 0.6 raises the predicted temporary layoff unemployment rate by about 0.5 percentage points, or one-third of the current average temporary layoff unemployment rate of 1.6 percent. Temporary layoff unemployment is more than twice as frequent among union members as among others between the ages of 25 and 55 who are in the experienced labor force. Unemployment insurance also has a correspondingly greater effect on that unemployment rate among union members: an increase in the *UI* benefit replacement ratio from 0.4 to 0.6 raises the predicted temporary layoff unemployment rate of union members by a full percentage point.

These estimates must be understood as subject to the biases and caveats discussed in Section I. It would clearly be desirable to repeat this research with *CPS* data for a more recent year. A reanalysis with data from the National Longitudinal Survey would be useful because the temporary layoff character of the unemployment spell could be defined *ex post*. It would be particularly valuable to extend the current data to include information on the experience rated tax of each individual's employer. More generally, it would be useful to reexamine the effect of *UI* on temporary layoffs by studying data on a sample of individual firms in a variety of states.

I have refrained throughout this paper from making any normative judgments about the effect of unemployment insurance on layoff unemployment. It is clear, however, that our current *UI* program does impose an efficiency loss by distorting the behavior of firms to lay off too many workers when demand falls rather than cutting prices or building inventories. The substantial rate of temporary layoff unemployment suggests that this efficiency loss may be quite large.

The redesign of unemployment insurance is a difficult problem because the unemployed include the job losers who must find new jobs as well as those on temporary layoff. For those who are changing jobs, the optimal insurance must balance providing protection from financial loss against the distortion to socially inefficient search.[32] For those who are on temporary layoff, it is sufficient to eliminate the *subsidy* element in *UI* by making each firm repay in taxes the full value of the benefits paid its employees and by making *UI* benefits subject to the same taxation as other compensa-

[31] These studies include Kathleen Classen; Ronald Ehrenberg and Ronald Oaxaca; Herbert Grubel and Dennis Maki; Arlene Holen and Stanley Horowitz, Charles Lininger; Marston. It should be clear that the only reliable studies of duration effects *exclude* temporary layoffs and combine data for individuals in different states or years. See Daniel Hamermesh and Finis Welch for discussions of this research.

[32] This point is discussed in more detail in my 1973 paper. Baily provides an excellent formal solution of this optimization problem.

tion.[33] The difficult problem arises because the full experience rating that is optimal for temporary layoffs is not optimal for permanent layoffs: it would inappropriately discourage new hiring and desirable layoffs.[34] The problem cannot be solved by a lower tax for those layoffs who are not rehired since that would distort the rehire decision and waste job-specific human capital. The optimum balancing of these considerations is a complex problem that requires more information than is currently available. A formal analysis of the problem would be valuable because it would indicate more precisely the type of information required and might provide new insights about the optimal design even before that information is collected.

As a practical solution, I believe that much could be gained by having full employer experience rating for the benefits paid during the first month of each spell of unemployment (or some other moderately short period). It would also be important to tax individuals on *UI* benefits in the same way as other compensation is taxed. This combination of reforms would eliminate most of the subsidy currently provided for short spells of temporary layoff unemployment without unduly discouraging either new hiring or permanent separations.[35]

[33] See the author (1976).

[34] Firms can often assess a worker's quality only after he has worked for the firm for a period of time. If layoffs of unsuitable workers are made very expensive by experience rating, firms will be reluctant to hire new workers and, when they make a hiring mistake, to discharge those who were inappropriately hired.

[35] The bias against new hiring could be reduced further by making the "one-month experience rating" provision apply only to workers with a minimum of, say, six months of experience with the firm.

REFERENCES

M. N. Baily, "Unemployment Insurance as Insurance for Workers," *Econometrica*, July 1977, *45*, 1043-63.

Joseph Becker, *Experience Rating in Unemployment Insurance*, Baltimore 1972.

B. R. Chiswick, "The Effect of Unemployment Compensation on a Seasonal Industry: Agriculture," mimeo., Hoover Instit., Stanford Univ. 1975.

K. Classen, "The Effects of Unemployment Insurance: Evidence from Pennsylvania," mimeo., Center for Naval Analysis, Washington 1975.

R. G. Ehrenberg and R. L. Oaxaca, "Unemployment Insurance, Duration of Unemployment, and Subsequent Wage Gain," *Amer. Econ. Rev.*, Dec. 1976, *66*, 754-66.

M. Feldstein, "Lowering the Permanent Rate of Unemployment," Study for the Joint Economic Comm., 93d Cong., 1st sess. 1973.

———, "Unemployment Compensation: Adverse Incentives and Distributional Anomalies," *Nat. Tax J.*, June 1974, *27*, 231-44.

———, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers*, Washington 1975, *3*, 725-45.

———, "Temporary Layoffs in the Theory of Unemployment," *J. Polit. Econ.*, Oct. 1976, *84*, 937-57.

R. B. Freeman, "Individual Mobility and Union Voice in the Labor Market," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 361-77.

H. G. Grubel and D. R. Maki, "The Effect of Unemployment Benefits on U.S. Unemployment Rates," *Weltwirtsch. Arch.*, submitted for publication.

D. Hamermesh, "Unemployment Insurance and Unemployment in the United States," mimeo., Univ. Illinois 1976.

A. Holen and S. Horowitz, "The Effect of Unemployment Insurance and Eligibility Enforcement on Unemployment," *J. Law Econ.*, Oct. 1974, *17*, 403-32.

Charles A. Lininger, Jr., *Unemployment Benefits and Duration*, Ann Arbor 1963.

S. T. Marston, "The Impact of Unemployment Insurance on Job Search," *Brooking Papers*, Washington 1975, *1*, 13-48.

J. L. Medoff, "Layoffs and Alternatives under Trade Unions in U.S. Manufacturing," mimeo., Harvard Univ. 1976.

Henri Theil, *Economic Forecasts and Policy*, Amsterdam 1961.

F. Welch, "What Have We Learned from

Empirical Studies of Unemployment Insurance?," mimeo., Univ. California-Los Angeles 1977.

U.S. Bureau of Labor Statistics, *Layoff, Recall, and Worksharing Procedures*, Bull. 1425-13, Washington 1972.

———, *Jobseeking Methods Used by American Workers*, Bull. 1886, Washington 1975.

———, *Current Population Survey (CPS)*, data reported in *Manpower Report of the President*, Washington, Mar. 1971.

Rural Wages, Labor Supply, and Land Reform: A Theoretical and Empirical Analysis

By MARK R. ROSENZWEIG*

Land reform is one of the most mentioned of the theoretical policy instruments discussed in the development literature, yet relatively little attention has been paid to the wage rate consequences of such a program, despite the fact that perhaps more than one-half of rural families in a developing country receive a substantial proportion of their income from wage earnings in agriculture.[1] One reason for this lacuna may be that the determination of wages and *family* labor supply in the agricultural sector of *LDC*s has also been somewhat neglected, particularly in the context of a heterogeneous labor force.[2] The subsistence or institutional wage models of W. Arthur Lewis, John Fei and Gustav Ranis, and Gerald Rodgers, for instance, offer no theory of how rural wage levels or differentials are set and thus provide little guidance on how wage rates would be affected by changes in land ownership patterns. More recently, Pranab Bardhan and T. N. Srinivasan, David Newbery, and Clive Bell and Pinhas Zusman, who formulate general equilibrium market or bargaining models determining endogenously the rental share paid by tenant sharecroppers, have assumed that agricultural wage rates are exogenous or determined only by nonagricultural factors. All of these models assume that rural labor is homogeneous.

Another reason why the potential wage impact of a land reform program may have received little attention is that models of "peasant" family behavior, such as those of A. K. Sen, Dipak Mazumdar, and Robert Mabro, typically embody two restrictive assumptions which would tend to make the equalization of landholdings appear wage augmenting, although this implication has never been formally derived. These assumptions are that 1) agriculture is "dualistic," with small-farm families facing lower shadow prices of labor (leisure) than large-farm landlords because of impediments to labor mobility and 2) agricultural households are "dichotomous"—"small" farmers employ family labor and maximize utility while "large" farms only utilize wage labor and maximize profits. Data from a 1970–71 all-India survey of over 5,000 households collected by the National Council of Applied Economic Research[3] indicate, however, that almost all cultivator households, large and small, participate actively in the labor market as either buyers or sellers of labor services, with almost 88 percent of households cultivating a gross-cropped area less than 1.5 acres utilizing some hired labor. Seventy-nine percent of these small farm households

*Associate professor, Yale University The research embodied in this paper was supported in part by a grant from the Ford and Rockefeller Foundations' Program in Support of Social Science and Legal Research on Population Policy and with funds from the U.S Agency for International Development under Order No. AID/otr-1432. Helpful suggestions and comments were provided by James L McCabe, Mark Gersovitz, an anonymous referee, the managing editor, and members of the Economic Growth Center, Yale University. I am also grateful to the members of the Research Program in Development Studies, Princeton University, for the use of their facilities. Research assistance was provided by James Devine, Anne Morgan, and Roberta Robson.

[1]Notable exceptions are R. Albert Berry, Mark Gersovitz, and M. A. Rahman All of these authors, however, employ geometric analyses with differing assumptions leading to wholly different "predictions" regarding wage effects. None consider the heterogeneity of agricultural labor, pay attention to questions of stability, or attempt to apply their models to data.

[2]Information on the differential impact of alternative agricultural policies, including land reform, on sex or age-specific wage rates is not only important in settling income distribution and equity issues but, as suggested in an article by the author and Robert Evenson, may have significant implications for population growth and schooling as well.

[3]For a more detailed discussion of these data, see the author.

had some family members who participated in the labor market with 55 percent reporting household members earning agricultural wages and working an average of 100 days in the market. Moreover, while almost 96 percent of the largest farms (30+ acres) hired labor, 85 percent also utilized family workers.

These data indicate the purchase of labor by almost all farms regardless of size and the extensive use of family labor by the largest farms. This suggests that the "dichotomization" of cultivating households by objective function would appear to be not only counterfactual but less useful than merely distinguishing large and small farms according to whether they are net importers or exporters of labor services, a distinction which identifies who benefits and who loses from a change in wage levels. Moreover, as will be shown below, when the dichotomy is dropped, the theoretical impact of a change in the distribution of landholdings on wage rates becomes ambiguous, with the possibility that wage rates may fall as a consequence of a land reform despite dualism and/or decreasing returns to scale in agricultural production.[4]

The primary objective of this paper is to formulate and test a general equilibrium model of rural wage determination. It embodies behavioral assumptions consistent with the labor mobility that characterizes the Indian agricultural labor market, and ascertains both theoretically and empirically the effects of a redistribution of land holdings on agricultural wage levels and sex/age wage differentials. In Section I, a competitive three-sector general equilibrium model of a dualistic agricultural labor market with two kinds of labor is formulated, and the stability and other properties of the equilibrium are described. In Section II, the impact of changes in agricultural labor supply on rural wages and wage differentials are assessed, and the effects of land reform on wages are derived and parameterized with respect to economies of scale, the extent of agricultural "dualism," differential income-leisure effects on large and small farms, and the relative disparity in landholdings. Section III contains an empirical analysis based on the theoretical framework in which I estimate the parameters of a six-equation simultaneous equations system describing the determination of rural wage rates and labor supply for males, females, and children based on aggregate Indian data. The results appear generally consistent with the competitive market model and indicate that rural wage levels and a measure of landholding inequality are negatively associated, but that an equalizing land redistribution would exacerbate agricultural wage differentials between males and females.

I. The Competitive Market Model and Properties of Equilibrium

To capture the essential features of rural agriculture and to maintain tractability, assume a labor market composed of two types of labor, male and female, and three agricultural households—a landless household and two households with different size plots, small and large, of quality-standardized land producing a homogeneous agricultural commodity. The market is initially assumed to be competitive so that all households are price takers, but wage rates are determined endogenously. There are, however, costs which vary with the labor time spent on the land owned by other households which are assumed to be borne entirely by workers.[5] Each household contains two persons, one of each labor type,

[4]In addition to these assumptions, Berry, who emphasizes the possibility of a wage decrease following a land redistribution, abstracts from labor-leisure choices in all households. Gersovitz in his nondualistic example assumes production is characterized by constant returns to scale and rules out negatively sloped labor supply curves. Rahman assumes constant returns to scale production and neglects labor-leisure choice.

[5]These costs are assumed to embody search and direct transportation costs and reflect the value of the disutility of off-farm work and the difficulties of distributing family income among members when some individuals are employed away from home. Off-farm labor costs per unit of labor time are assumed to be exogenous and invariant with respect to time worked Relaxation of the latter assumption introduces considerable complexity into the analysis.

each owning a unit of labor time. The two types of labor are imperfect substitutes in agricultural production, but labor of each type from different households is perfectly substitutable.[6] No labor is sold outside the agricultural sector.

The landless household supplies $l_{fm}^N = 1 - l_M^N$ and $l_{fW}^N = 1 - l_W^N$ amounts of labor to the market, where l_M^N and l_W^N are the quantities of leisure time of the husband and wife in the landless household. Total consumption of the landless family, assuming no saving and a unit price for the composite consumption commodity, is thus

$$(1) \qquad X^N = l_{fM}^N \Pi_M^N + l_{fW}^N \Pi_M^N$$

where $\Pi_K^N = W_K - \rho_K (K = M, W)$; W_K are the market wages paid to (hired) male and female labor, and ρ_K is the cost per unit of labor time supplied to the market, assumed to be exogenous.

The small-farm household owns A^S units of land and is by definition a net exporter of the labor services of both the husband and wife. The large-farm household owns θA^S units of land, where θ is a scalar chosen such that the household is an importer of labor. Denoting L_M^i and L_W^i, $i = S, L$, as the total amounts of male and female labor utilized on the land owned by each landowning household, the quantities of male and female labor supplied (exported) to the market by the small household, λ_M^S and λ_W^S, and the amounts of labor hired (imported) by the large landowning family λ_M^L and λ_W^L are given by

$$(2) \qquad \lambda_K^S = l_{fK}^S - L_K^S > 0$$

$$(3) \qquad \lambda_K^L = L_K^L - l_{fK}^L > 0 \qquad K = M, W$$

where l_{fK}^i is the total work time of family member K on the farm of size i.

The quantities consumed by the landowning households, X^S and X^L, are thus

$$(4) \quad X^i = F(L_M^i, L_W^i, \theta^j A^S) + (-1)^j \lambda_M^i \Pi_M^i + (-1)^j \lambda_W^i \Pi_W^i \qquad i = S, L$$
$$j = 0 \text{ for } i = S;\ 1 \text{ for } i = L$$

[6]Also assume that the land market is imperfect, such that the distribution of land is fixed, ignore other agricultural inputs, and abstract from uncertainty, seasonality, and land tenure considerations.

where $\Pi_K^L = W_K$, $\Pi_K^S = W_K - \rho_k$ and F is a twice continuously differentiable strictly concave production function with positive cross partials, which may exhibit either decreasing, constant, or increasing returns to scale.

Each of the three households maximizes an identical twice-differentiable family utility function given by (5), with respect to the consumption commodity X^i and the leisure of the two household members, each of which is assumed to be noninferior, subject to the relevant budget constraints in (1) and (4).

$$(5) \quad U = U(X^i, l_M^i, l_W^i) \qquad i = N, S, L$$

If only interior solutions are considered, the necessary conditions for each household, in addition to those implied by the budget constraints, are given by equations (6) through (8):

$$(6) \qquad U_X^i - \Psi^i = 0 \qquad i = N, S, L$$

$$(7) \quad U_K^i - \Psi^i \Pi_K^i = 0 \qquad i = N, S, L$$

$$(8) \qquad F_K^i - \Pi_K^i = 0 \qquad i = S, L$$

where U_K^i is the partial derivative of (5) with respect to l_K^i in households of type i, F_K^i is the marginal product of L_K^i in farm households of type i, and ψ^i is the Lagrangean multiplier for household i.

Equations (7) and (8) give the standard utility- and profit-maximizing results describing the optimal quantities of leisure and total labor use, if any, for each household. With $\rho_K > 0$, the market is dualistic in the sense that small landowning households utilize more labor per acre than large landowners because of the differential shadow prices of labor: $F_K^S < W_K$, $F_K^L = W_K$. Each member of the small landowning household allocates his (her) labor on the family's land up to the point where the value of his (her) marginal product just equals the net wage he (she) receives in the market, $W_K - \rho_K$. Members of the large landowning households devote all their *work* time to their own land and hire each type of labor up to the point at which the marginal value product of that labor type is equal to the appropriate market wage W_K.

To derive the partial-equilibrium comparative static properties for the three households, first write the matrix:

$$\beta^i = \begin{bmatrix} U^i_{XX} & U^i_{XM} & U_{XW} & -1 \\ U^i_{XM} & U^i_{MM} & U^i_{WM} & -\Pi^i_M \\ U^i_{XW} & U^i_{MW} & U^i_{WW} & -\Pi^i_w \\ -1 & -\Pi^i_M & -\Pi^i_w & 0 \end{bmatrix}$$

$$i = N, S, L$$

Differentiating equations (1), (6), and (7) for $i = N$, we get equation (9), where β^N is thus the bordered Hessian matrix for the landless household. Denoting the determinant of β^i as ϕ^i and the cofactor of row r and column c of β^i as ϕ^i_{rc}, we obtain the standard Slutsky equations for the landless household's labor supply:

$$(10)\quad \frac{dl^N_{fK}}{dW_K} = -\Psi^N\frac{\phi^N_{nn}}{\phi^N} + l^N_{fK}\frac{\phi^N_{4k}}{\phi^N} = \sigma^N_{KK} - l^N_{fK}\sigma^N_K \qquad K = M, n = 2; \quad K = W, n = 3$$

$$(11)\quad \frac{dl^N_{fK}}{dW_h} = -\Psi^N\frac{\phi^N_{23}}{\phi^N} + l^N_{fh}\frac{\phi^N_{4n}}{\phi^N} = \sigma^N_{Kh} - l^N_{fh}\sigma^N_K \qquad \text{for } K \neq h$$

where σ^N_K is the income effect on leisure for family member K.

Second-order conditions constrain the first term in equation (10), the compensated substitution effect σ^N_{KK}, to be positive, since $\phi^N < 0$ and $\phi^N_{nn} > 0$. The normality assumption, however, implies that the second term, containing the income effect on leisure σ^N_K, is negative so that equation (10) is

$$(9)\quad [\beta^N]\begin{bmatrix} dX \\ dl^N_M \\ dl^N_W \\ d\Psi^N \end{bmatrix} = \begin{bmatrix} 0 \\ \Psi^N \; dW_M - \Psi^N \; dp_M \\ \Psi^N \; dW_W - \Psi^N \; dp_W \\ -l^N_{fM} \; dW_M - l^N_{fW} \; dW_w + l^N_{fM} \; dp_M + l^N_{fW} \; dp_W \end{bmatrix}$$

$$(12)\quad \begin{bmatrix} & & & 0 & 0 & -1 \\ & \phi_{44} & & 0 & 0 & -\Pi^i_M \\ & & & 0 & 0 & -\Pi^i_W \\ 0 & 0 & 0 & F^i_{MM} & F^i_{MW} & 0 \\ 0 & 0 & 0 & F^i_{MW} & F^i_{WW} & 0 \\ -1 & -\Pi^i_M & -\Pi^i_W & 0 & 0 & \end{bmatrix}\begin{bmatrix} dX^i \\ dl^i_M \\ dl^i_W \\ dL^i_M \\ dL^i_W \\ d\Psi^i \end{bmatrix} =$$

$$= \begin{bmatrix} -\Psi^i dW_M \\ -\Psi^i dW_W \\ -dW_M \quad -F_{L_M A^i} \quad dA^i \\ -dW_W \quad -F_{l_W A^i} \quad dA^i \\ (L^i_M - l^i_{fM}) \; dW_M + (L^i_W - l^i_{fW}) \; dW_W - F_{A^i} dA^i \end{bmatrix}$$

consistent with either a backward-bending or positively sloped supply curve for landless laborers of either sex. The sign of (11) depends on whether the leisure time of the husband and wife are complements or substitutes, being negative if the leisure time of spouses are substitutes. Total differentiation of equations (4) and (6) through (8) for $i = S, L$ yields equation (12), where ϕ_{44} is the 3 x 3 cofactor in β^i. Noting that β^i is therefore the second bordered principal minor of the bordered Hessian matrix in (12) and must be negative, we obtain the following results for the two landowning households, employing Cramer's rule:

$$(13)\quad \frac{dl^i_{fK}}{dW_K} = -\Psi^i \frac{\phi^i_{nn}}{\phi^i} + (L^i_K - l^i_{fK})\frac{\phi^i_{4n}}{\phi^i} = \sigma^i_{KK} - (L^i_K - l^i_{fK})\sigma^i_K \qquad n = 2, 3$$

$$(14)\quad \frac{dl^i_{fK}}{dW_h} = -\Psi^i \frac{\phi^i_{23}}{\phi^i} + (L^i_h - l^i_{fh})\frac{\phi^i_{4n}}{\phi^i} = \sigma^i_{Kh} - (L^i_h - l^i_{fh})\sigma^i_K$$

$$(15)\quad \frac{dL^i_K}{dW_h} = -\frac{F^i_{Kh}}{\Delta^i} \begin{array}{l} < 0 \text{ for } K = h \\ > 0 \text{ for } K \neq h \end{array}$$

$$(16)\quad \frac{dl^i_{fK}}{dA^i} = F_{A^i}\frac{\phi^i_{4n}}{\phi^i} = -F_{A^i}\sigma^i_K < 0$$

$$(17)\quad \frac{dL^i_K}{dA^i} = \frac{F^i_{hA}F^i_{MW} - F^i_{KA}F^i_{hh}}{\Delta^i}$$

where $\Delta^i = F^i_{WW}F^i_{MM} - (F^i_{MW})^2 > 0$

Equations (13) and (14), which give the own- and cross-wage effects on the total supply of work time for each household member in the landowning households, indicate that the substitution and income effects in those households are qualitatively similar to those of the landless households. They are identical if the labor market is nondualistic and competitive ($\Pi^N_K = \Pi^S_K = W_K$) and if the utility function in (5) is homothetic. However, unlike landless laborers and small landowners who are labor exporters, the uncompensated own-wage effect on total (family) labor supply in labor-importing farms is unambiguously positive, since a wage rise must lower net income for these households.

Equations (15) and (17), giving (own and cross) effects of a rise in wage rates and landholdings on total labor usage on the landowning farms, indicate that if competitive conditions prevail, the partial equilibrium changes in the allocation of total farm labor will be identical whether or not (some) households maximize utility or profits. However, as will be shown below, the assumption that large landowners maximize utility and utilize family labor has consequences for the allocation of market (nonfamily) labor, and thus for the levels of the equilibrium wage rates and the stability of the rural labor markets, which are functions of *market* supply and demand curves only.

The relationship between the supply of off-farm labor of type K from small farms and changes in wage rates, from (13), (14), and (15), is expressed in (18):

$$(18)\quad \frac{d\lambda^S_K}{dW_h} = \left[\sigma^S_{Kh} - \frac{F^S_{Kh}}{\Delta^S}\right] - \lambda^S_h\sigma^S_K$$

While for own-wage effects ($K, h = M$ or W) the terms in brackets, the own-compensated substitution effect and the negative of the labor usage effect, must be greater than zero, (18) may be of either sign because of the positive income effect on leisure σ^S_K.

For the labor-importing utility-maximizing farms, the own- and cross-wage effects on the quantity of labor of sex K hired, λ^L_K, is given by

$$(19)\quad \frac{d\lambda^L_K}{dW_h} = \left[\frac{F^L_{Kh}}{\Delta^L} - \sigma^L_{Kh}\right] - \lambda^L_h\sigma^L_K$$

Since the demand for all labor of type K to be used in agricultural production falls and the quantity of labor supplied by family members of sex K increases when W_K rises, the demand for hired labor must decline in response to a wage rise.

The effects of an exogenous increase in household landholdings on off-farm labor supply (small farms) and on the demand for hired labor on type K (large farms) depend

also on both production and income-leisure effects, but are of *un*ambiguous signs.

$$(20)\quad \frac{d\lambda_K^S}{dA} = -F_A^S \sigma_K^S - \left[\frac{F_{hA}^S F_{Kh}^S - F_{KA}^S F_{hh}^S}{\Delta^S}\right] < 0$$

$$(21)\quad \frac{d\lambda_K^L}{dA} = F_A^L \sigma_K^L + \left[\frac{F_{hA}^L F_{Kh}^L - F_{Ka}^L F_{hh}^L}{\Delta^L}\right] > 0$$

$$K = M, h = W$$
$$K = W, h = M$$

It is seen that an increase in the size of labor-exporting farms will reduce their supply of labor to other farms; an increase in the holdings of labor-importing households will increase the demand for hired labor because of reinforcing production and income-leisure effects, even if scale diseconomies (which do not violate second-order conditions) exist.

Labor market equilibrium is characterized by equations (1), (4), and (6)–(8), as well as equilibrium conditions (22):

$$(22)\quad l_{fK}^N + \lambda_K^S = \lambda_K^L \quad K = M, W$$

A necessary condition for (Hicksian) multi-market stability in the market for hired agricultural labor, from equations (13), (18), and (19), is that

$$(23)\quad \frac{d(\lambda_K^L - \lambda_K^S - l_{fK}^N)}{dW_K} = \sum_{i=S,L} \frac{F_{KK}^i}{\Delta^i} - \sum_{i=S,L,N} \sigma_{KK}^i - \lambda_K^L \sigma_K^L + l_{fK}^N \sigma_K^N + \lambda_K^S \sigma_K^S < 0$$

The assumptions imposed in the analysis so far do not insure that condition (23) will be met; it is thus possible that with sufficiently negative sloped *market* supply curves of agricultural labor, the market equilibrium will not be stable. However, the likelihood that instablity is the major reason for the existence of institutional (i.e., nonmarket determined) wages is low: the presence of labor-hiring institutions (large landowners) which maximize utility and employ family labor, as in India, as well as the existence of labor-supplying households whose members both work their own land and offer labor services to the market, makes the fulfillment of the static stability conditions more likely in the context of Indian agriculture than in developed country (modern sector) labor markets. In the latter, where employers of hired labor are profit maximizers, and household members who supply labor do not participate in household income production, three negative terms tending toward stability, F_{KK}^S/Δ^S, $-\sigma_{KK}^S$, and $-\lambda^L\sigma_K^L$, would not appear in (23). Moreover, because of the participation of family members in agricultural production on labor-importing farms, the stability condition *must* be satisified if the utility function is homothetic (and $\rho = 0$) since the last three terms in (23) vanish; i.e., if $\sigma_K^L = \sigma_K^N = \sigma_K^S$, then from (22), $-\lambda_K^L\sigma_K^L + l_{fK}^N\sigma_K^S + \lambda_K^S\sigma_K^S = 0$.

II. General Equilibrium Comparative Statics

Assuming a unique stable equilibrium we can ascertain the effects of a change in landholdings A^i (or any other exogenous variable hypothesized to influence supply behavior) on the wage rates of the two types of labor by totally differentiating equations (1), (4), (6)–(8), and (22), and solving for dW_M and dW_W. First we show that the attenuation of factors inhibiting the participation in market work of only one group, say, females, will lower agricultural wage rates generally but will not necessarily result in wider intergroup wage differentials.[7] To see this let R be an environmental characteristic such that dl_{fW}^N/dR, $d\lambda_W^S/dR < 0$; dl_{fM}^N/dR, $d\lambda_M^S/dR$, $d\lambda_K^L/dR = 0$, then for a small change in R around equilibrium the effect on male and female agricultural wage rates can be written in terms of the partial equilibrium comparative static results, where

[7]Ester Boserup hypothesizes that high levels of female participation rates are associated with greater disparities in male-female wage differentials.

$\epsilon^i_{KX} = d\lambda^i_K/dX$ and $\mu^i_{Kh} = d\lambda^i_K/dW_h$:

$$(24)\quad \frac{dW_W}{dR} = \left[\frac{dl^N_{fW}/dR + d\lambda^S_W/dR}{(\mu^L_{WW} - \mu^S_{WW} - \mu^N_{WW})}\right] \cdot \Omega^{-1} > 0$$

$$(25)\quad \frac{dW_M}{dR} = -\left[\frac{dl^N_{fW}/dR + d\lambda^S_W/dR}{(\mu^L_{WW} - \mu^S_{WW} - \mu^N_{NW})}\right] \cdot \left[\frac{(\mu^L_{WM} - \mu^S_{WM} - \mu^N_{WM})}{(\mu^L_{MM} - \mu^S_{MM} - \mu^N_{MM})}\right] \Omega^{-1} > 0$$

where $\Omega = 1 - \dfrac{(\mu^L_{MW} - \mu^S_{MW} - \mu^N_{MW})(\mu^L_{WM} - \mu^S_{WM} - \mu^N_{WM})}{(\mu^L_{MM} - \mu^S_{MM} - \mu^N_{MM})(\mu^L_{WW} - \mu^S_{WW} - \mu^N_{WW})}$

To sign (24) and (25) note that the assumptions of strict concavity in production and the second-order conditions require that $\Omega > 0$ and that if the equilibrium is dynamically stable, from (23), $(\mu^L_{KK} - \mu^S_{KK} - \mu^N_{KK}) < 0$ and $(\mu^L_{Kh} - \mu^S_{Kh} - \mu^N_{Kh}) > 0$.[8] Thus, expressions (24) and (25) must be greater than zero; an increase in female market participation must reduce male and female wage rates, the magnitude of the effects being positively related to the sensitivity of female labor supply to changes in R and negatively to the sensitivity of the demand and supply of hired labor to "own" changes in agricultural wage rates. However, the change in the wage rate differential given by (26) cannot be predicted:

$$(26)\quad \frac{d(W_M - W_W)}{dR} = \left[\frac{dl^N_{fN}/dR + d\lambda^S_W/dR}{(\mu^L_{WW} - \mu^S_{WW} - \mu^N_{WW})}\right] \cdot \left[\frac{(\mu^L_{WM} - \mu^S_{WM} - \mu^N_{WM})}{(\mu^L_{MM} - \mu^S_{MM} - \mu^N_{MM})} - 1\right] \Omega^{-1}$$

Using these results, we derive the effect of a redistribution of land (without compensation for the transfer to wealth) from large to small landowners on wage rates in the general equilibrium system by solving for the effects of an increase in A^S on W_W and W_M under the side condition that total landholdings $A^T = A^S(1 + \theta)$ remain constant:

$$(27)\quad \frac{dW_K}{dA^S} = \left[\frac{\epsilon^L_{KA} + \epsilon^S_{KA}}{(\mu^L_{KK} - \mu^S_{KK} - \mu^N_{KK})} - \frac{\epsilon^L_{hA} + \epsilon^S_{hA}}{(\mu^L_{KK} - \mu^S_{KK} - \mu^N_{KK})} \cdot \frac{(\mu^L_{Kh} - \mu^S_{Kh} - \mu^N_{Kh})}{(\mu^L_{hh} - \mu^S_{hh} - \mu^N_{hh})}\right] \Omega^{-1}$$

Assuming that the direct effect, the first bracketed term, dominates, the sign of (27) depends on the sign of $\epsilon^L_{KA} + \epsilon^S_{KA}$, so that from (20) and (21):

$$(28)\quad \frac{dW_K}{dA^S} \gtreqless 0$$

as $$\frac{F^S_{hA}F^S_{Kh} - F^S_{KA}F^S_{hh}}{\Delta^S} - \frac{F^L_{hA}F^L_{KA} - F^L_{KA}F^L_{hh}}{\Delta^L} + F^S_A\sigma^S_K - F^L_A\sigma^L_K \gtreqless 0$$

Thus whether or not a land reform program without compensation[9] increases or decreases the wage rates for laborers of type (sex) K depends on the properties of the production function and the differences in income-leisure relationships for individuals of sex K and the marginal product of land on small and large farms. To parameterize these relationships assume that the production function is Cobb-Doublas,[10] such that

$$F = Q^j = (L^j_m)^{\beta_1}(L^j_w)^{\beta_2}(\theta^j A)^{\beta_3}$$

and $\beta_1 + \beta_2 < 1$. Expression (28) can then be rewritten as:

[8]The second inequality embodies the condition that wage laborers of each type are gross substitutes, which guarantees dynamic local stability for all speeds of adjustment. See Kenneth Arrow, Harold Block, and Leonid Hurwicz.

[9]The degree of compensation can be easily introduced into the analysis as a parameter. As long as compensation is not complete, so that both the recipients and the donors of land experience opposite changes in real wealth (apart from indirect wage effects), income-leisure effects will be relevant.

[10]Bardhan, fitting a number of alternative functional forms to Indian production data, could not reject the Cobb-Douglas function. However, the conclusions derived from the model are not dependent on the functional form chosen as long as the function is well behaved.

$$(29) \quad \frac{dW_K}{dA^S} \gtreqless 0$$

$$\text{as } \gamma \frac{L_K^S}{A^S}\Big[i - \theta^{(\gamma-1)}\Big(\frac{W_K - \rho_K}{W_K}\Big)^{\frac{1-\beta_h}{1-\beta_1-\beta_2}} \cdot \Big(\frac{W_h - \rho_h}{W_h}\Big)^{\frac{\beta_h}{1-\beta_1-\beta_2}}\Big] - \beta_3 \frac{Q^S}{A^S}\Big[\sigma_K^L \theta^{(\gamma-1)} \cdot \Big(\frac{W_K - \rho_K}{W_K}\Big)^{\frac{1-\beta_h}{1-\beta_1-\beta_2}} \cdot \Big(\frac{W_h - \rho_h}{W_h}\Big)^{\frac{\beta_h}{1-\beta_1-\beta_2}} - \sigma_K^S\Big] \gtreqless 0$$

where $\gamma = \beta_3/(1 - \beta_1 - \beta_2)$

The following conclusions emerge:

1) With no factor distortions ($\rho = 0$), linear homogeneity ($\gamma = 1$), increasing returns to scale ($\gamma > 1$), or decreasing returns to scale ($\gamma < 1$) are each neither sufficient nor necessary for land redistribution to be wage neutral ($dW_k/dA^S = 0$), wage augmenting, or wage decreasing because of income-leisure effects. If the production function is linear homogeneous, moreover, the differences between income-leisure effects in small- and large-farm households will uniquely determine the direction of the wage effect, assuming compensation if any is not complete. Since that differential may be of opposite sign for males and females, it is possible that land reform could raise wage rates for one group while lowering them for another.

2) In the special case considered by Gersovitz, Mabro, and others, in which the production function is linear homogeneous and large farms are owned by profit-maximizing absentee landlords (no employment of family labor so $\sigma_K^L = 0$), wage rates of men and women will rise unambiguously; the magnitude of the rise, from (27), being a negative function of the sensitivity of the demand and supply of hired labor to wage-rate changes and a positive function of the magnitude of the income-leisure effects on small-farm households. In this case, the wage group benefiting most from the land reform will be that which has the greatest income elasticity of leisure and the most inelastic market demand and supply curves.

3) Sufficient but not necessary conditions for land reform to be wage neutral under competitive conditions (with $\rho = 0$) are that the production function be linear homogeneous *and* the utility function be homothetic, neither assumption by itself is necessary or sufficient.

4) Dualism in agriculture does not necessarily imply that land reform will increase rural wages. Moreover, rural wages can rise after a land reform without factor distortions. However, the greater the costs to workers of off-farm employment, the more likely will wages rise as a result of a land redistribution. To see this, differentiate (29) with respect to ρ_K, noting that $\beta_h < 1$.

$$(30) \quad \Big(\frac{1-\beta_h}{1-\beta_1-\beta_2}\Big)\Big(\frac{W_K - \rho_K}{W_K}\Big)^{\frac{1-\beta_h}{1-\beta_1-\beta_2}-1} \cdot \Big(\frac{W_h - \rho_h}{W_h}\Big)^{\frac{\beta_h}{1-\beta_1-\beta_2}} \theta^{(\gamma-1)} \cdot \Big[\gamma \frac{L_K^S}{A^S} + \sigma_K^L \beta_3 \frac{Q^S}{A^S}\Big] > 0$$

III. Empirical Analysis

A. *Variables and Reduced-Form Estimates*

The principal implication of the preceding theoretical analysis is that the direct impact of a land redistribution program on agricultural wage rates is indeterminate. Moreover, as was demonstrated, data pertaining to scale economies[11] and the labor-supply elasticities of landless and landowning households would provide only indirect

[11]The evidence from Indian data is mixed. Stanislaw Wellisz, using aggregate pooled time-series data from Andhra Pradesh, concluded that agricultural production was characterized by increasing returns to scale. Bardhan, however, found evidence of decreasing returns to scale in paddy agriculture and constant returns to scale in wheat growing areas based on individual farm data from seven districts.

evidence on the consequences of land reform policy and would not, in any event, indicate the quantitative magnitude of its impact on rural wages. In this section a more direct approach is adopted, utilizing the general equilibrium market framework and aggregate data from India to estimate the direct *ceteris paribus* relationship if any between the size distribution of landholdings and the wage rates of adult males, adult females, and children in the agricultural sector, thereby obtaining a quantitative estimate of the potential wage impact of a land-redistribution program. To assess the applicability of the general equilibrium model, a number of the verifiable predictions of the model regarding the relationships between the distribution of land, rural wage rates, and the supply of and demand for hired labor are also tested.

Let us first estimate a set of six reduced-form equations in which the levels of the agricultural wage rates of adult males and females and children and the number of wage laborers per household in each sex group are the dependent variables, using data pertaining to the rural populations in 159 Indian districts, 1960-61.[12] The maintained hypothesis motivating the empirical analysis is that interdistrict labor mobility in India is sufficiently low such that district-level characteristics are the important determinants of rural district wage rates and market labor supply.

$$(31) \qquad W_{kj} = a_{k0} + \sum_{i=1}^{16} a_{ki}X_{ij} + u_{kj}$$

$$(32) \qquad \lambda_{kj} = b_{k0} + \sum_{i=1}^{16} b_{ki}X_{ij} + v_{kj}$$

$$k = M, W, C; j = 1 \ldots 159$$

Each of the six equations, (31) and (32), contains an identical set of exogenous explanatory variables X_i, listed and defined in Table 1, which also provides means and standard deviations. The set of regressors includes three variables characterizing the size distribution of land: the mean holdings of landowners X_1, the proportion of households in rural areas without land X_2, and a measure of landholding inequality among landowners X_3, the Kuznets ratio, given by (33):[13]

$$(33) \qquad X_3 = \sum_{i=1}^{12} \left| \frac{P_{ij}}{P_j} - \frac{A_{ij}}{A_j} \right|$$

where P_j = total number of landowning households in district j

P_{ij} = number of landowning households in interval i in district j

A_j = total landholdings (acres) in district j

A_{ij} = landholdings in interval i in district j

Other variables included in the set of exogenous regressors standardize for differences in land-augmenting factors, such as rainfall and irrigation, and represent other rural population characteristics and institutions as well as nonagricultural demand factors.[14] These latter variables are assumed to draw labor out of the agricultural sector and can be shown, based on straightforward manipulations of the model described in Sections I and II, to raise agricultural wage levels.

With respect to the economic characteristics of the agricultural sector, the model, while agnostic in terms of the signs of the distribution variable coefficients in the reduced-form wage equations, does make the following predictions:

1) The coefficients of the mean landholdings, irrigation, and rainfall variables should display positive signs in all wage equations. An increase in average landholdings or land-augmenting factors per

[12]The criterion for district inclusion is that wage rates be reported for at least one month of the year for all three groups. The districts selected are thus not necessarily representative of India as a whole, although they cover a wide geographic area.

[13]This measure was chosen for computational ease and because of its well-known property of being sensitive to changes occurring at the tails of the distribution, where a land reform program is likely to operate. Experimentation with alternative distributional parameters, such as the *log* variance and the Gini coefficient, on a subset of districts produced insignificant changes in results.

[14]For a discussion of the role of the "control" variables, see the author.

TABLE 1—VARIABLE DEFINITIONS, MEANS, AND STANDARD DEVIATIONS, 159 INDIAN DISTRICTS,[a] 1960–61

Variable	Definition	Mean	Standard Deviation
W_M	Daily wage in rupees for male field labor (sowers, reapers, weeders, ploughers)	1.52	0.43
W_W	Daily wage in rupees for female field labor (sowers, reapers, weeders, ploughers)	1.11	0.37
W_C	Daily wage in rupees for child field labor and herding	0.85	0.37
λ_M	Percentage of males per household age 15–59 working at least one hour per day as hired agricultural laborers	23.4	11.2
λ_W	Percentage of females per household age 15–59 working at least one hour per day as hired agricultural laborers	22.0	14.4
λ_C	Percentage of children per household age 5–14 working at least one hour per day as hired agricultural laborers	5.75	3.98
X_1	Average land owned per landowning household	12.4	10.3
X_2	Percentage of households without land	34.9	13.1
X_3	Kuznets ratio of landholding inequality	81.7	16.3
X_4	Percentage of males 15–59 with primary education	12.7	9.27
X_5	Percentage of females 15–59 with primary education	3.34	4.11
X_6	Percentage of males 15–59 with secondary education	2.44	2.50
X_7	Percentage of females 15–59 with secondary education	0.27	0.68
X_8	Percentage of population in scheduled tribes	12.8	6.32
X_9	Percentage of cultivated acres irrigated	12.8	17.4
X_{10}	Average normal rainfall per year in cm	302.2	594.2
X_{11}	Dummy = 1 if at least one plantation in district	0.10	
X_{12}	Factories and workshops per household	0.17	0.18
X_{13}	Percentage of factories and workshops employing 5+ persons	3.9	4.0
X_{14}	Percentage of factories and workshops using power	20.5	19.2
X_{15}	Proportion of population living in urban areas	0.17	0.11
X_{16}	Percentage of population Moslem	33.2	66.6

Sources: See Appendix.
[a]States in data sample: Andhra Pradesh, Assam, Bihar, Gujurat, Kerala, Madhya Pradesh, Madras (Tamil Nadu), Maharashtra, Mysore, Orissa, Punjab (and Haryana), and Uttar Pradesh.

household, controlling for the distribution of land among landholders and the proportion of landless households, from (20) and (21), would increase the demand for hired labor on labor-importing farms and decrease the supply of off-farm work from labor-exporting households.

2) The proportion of households without land should be positively associated with the employment of wage laborers and negatively correlated with the wage levels of all sex-age groups. Landless households supply more labor to the market than those households owning land.

The ordinary least squares (OLS) reduced-form parameter estimates are presented in Table 2. The set of district-level variables explains approximately 35–47 percent (adjusted for degrees of freedom) of the inter-district variation in rural male, female, and child wage rates, with the highest explanatory power being obtained for adult male wages. The same variables account for 53–60 percent of the variation across districts in wage laborers per household for the three sex-age groups.

The results are generally consistent with the predictions of the model: the coefficients

TABLE 2—UNRESTRICTED REDUCED-FORM *OLS* COEFFICIENT ESTIMATES, INDIAN DISTRICTS, 1960–61

Independent Variable	Dependent Variable					
	W_M	W_W	W_C	λ_M	λ_W	λ_C
X_1	.0187	.0136	.0054	−.0594	−.124	−.0030
	(4.66)	(3.81)	(1.40)	(0.61)	(1.05)	(0.09)
X_2	−.0018	−.0018	−.0004	.380	.405	.0906
	(0 53)	(0 59)	(0.13)	(4.57)	(4.04)	(3.32)
X_3	−.0133	− 0101	− 0062	.355	430	.120
	(6 39)	(5.42)	(3.12)	(6.96)	(6.99)	(7.22)
X_4	0140	0091	.0099	.219	.133	−.0483
	(2.44)	(1 79)	(1.81)	(1.56)	(0.78)	(1.05)
X_5	− 0019	0064	−.0040	− 152	−.282	− 0260
	(0 14)	(0 51)	(0.30)	(0 44)	(0.68)	(0.23)
X_6	0095	0056	.0002	−.262	−.189	−.118
	(0 64)	(0 43)	(0 13)	(0.72)	(0 43)	(1.00)
X_7	.0793	.0280	.0792	−1.83	−5.58	−1.58
	(1 06)	(0 42)	(1 10)	(1.00)	(2.53)	(2.63)
X_8	0092	0149	0166	− 243	−.522	− 128
	(1 74)	(3 18)	(3 29)	(1 88)	(3 36)	(3.04)
X_9	.0059	0033	0006	0413	−.0166	−.0052
	(2.69)	(1.66)	(0 31)	(0.77)	(0.25)	(0.30)
X_{10}	.0003	0002	0001	− 0006	0024	.0004
	(3 20)	(2.69)	(1 18)	(0 32)	(1 04)	(0.67)
X_{11}	−.196	− 185	− 148	−1.23	−4.36	−1.44
	(1 51)	(1 60)	(1 19)	(0 39)	(1 14)	(1 39)
X_{12}	−.0027	110	0495	−4 42	−7 37	−2.88
	(0 02)	(0 72)	(0 30)	(1 05)	(1 45)	(2.09)
X_{13}	− 0039	− 0006	.0010	− 210	− 260	− 120
	(0 30)	(0 07)	(0.10)	(0.85)	(0.87)	(1.49)
X_{14}	0050	0067	0064	− 0213	−.176	−.0192
	(2 75)	(4 16)	(3.68)	(0 48)	(3 32)	(1.33)
X_{15}	501	.514	318	−14 13	−10.78	−3.88
	(1 81)	(2 09)	(1 20)	(2.09)	(1 32)	(1 75)
X_{16}	− 0004	.0012	0019	0376	.0023	−.0094
	(0.59)	(1 94)	(2 72)	(2 14)	(0 11)	(1.63)
Constant	2 31	1 64	0.99	−26 82	−21 11	−5.74
	(7 31)	(5 84)	(3 28)	(3 47)	(2 26)	(2.27)
S.E E	331	295	318	8 09	9.77	2.65
$\bar{R}^2$	465	424	349	.534	.587	.603

Note The *t*-values are shown in parentheses. Number of districts = 159

of the mean landholding, rainfall, and irrigation variables display the predicted signs in the wage equations, the coefficients being statistically significant in all but the child wage equation, while the proportion of landless households as expected is positively associated with the proportion of laborers in agricultural employment. Agricultural wages appear to be influenced as well by factors outside the agricultural sector, with nine of the twelve coefficients of these variables displaying the expected signs.

Most importantly, the coefficients of the land distribution variable, strongly significant in all equations, suggest that wage rates of men, women, and children are lower and market employment higher where the distribution of land is most unequal. These results thus suggest that a redistribution of land among landholders which reduced landholding inequality would raise agricultural wages in India. The differences in the land distribution coefficients in the three wage equations, statistically significant at the 1 percent level, however, sug-

gest that reductions in land inequality would exacerbate arithmetic sex-age wage differentials in rural areas.

B. *Structural Estimates*

To fully test the market wage model I estimate structural demand and supply equations for hired labor, described by (34) and (35):

$$(34)\quad W_{kj} = \alpha_{ko} + \sum_{h} \alpha_{kh}\lambda_{hj} + \sum_{i=1}^{11} \alpha_{ki}x_{ij} + \epsilon_{kj} \qquad k,h = M,W,C$$

$$(35)\quad \lambda_{kj} = \gamma_{ko} + \gamma_k W_k + \sum_{i=1}^{8} \gamma_{ki}X_{ij} + \sum_{i=12}^{16} \gamma_{ki}X_{ij} + e_{kj}$$

where the ϵ_{kj}, e_{kj} are stochastic error terms.

The demand equations in (34) have been specified with the wage rate as the dependent variable so that the direct influence of labor-supply changes on wage rates can be more easily tested. If wage rates are influenced by shifts in supply and demand as implied by the theoretical analysis, all the α_{kh} coefficients should display negative signs, since from (24) and (25) an increase in the quantity of labor of type K must have negative own- and cross-wage effects in equilibrium. The theoretical analysis also suggests that mean landholding size, the extent of irrigation, and the quantity of rainfall should be positively associated with the demand for hired labor. Moreover, the demand for *wage* labor should be highest in areas where landholding inequality is greatest *if the labor market is competitive*, even if scale diseconomies exist, since where the distribution of landholdings is more unequal more land is held by labor-importing farm households.

With respect to the supply equations, the own-wage effects on labor supply are theoretically ambiguous as was shown;[15] however, the model suggests that the proportion of landless households and the degree of landholding inequality should be positively associated with the supply of market labor from (20), since an increase in the magnitude of these variables is equivalent to a reduction in the landholdings of labor-exporting households. Similarly, an increase in mean landholdings should decrease the supply of agricultural market workers per household as should increases in the nonagricultural labor demand variables.

Because the residual correlation matrix obtained from the estimation of (34) and (35) by two-stage least squares indicated significant correlations of residuals across equations, the system of market demand and supply equations was estimated using full information maximum likelihood (*FIML*) to capture potential efficiency gains.[16] The estimate *FIML* coefficients are displayed in Table 3.

The structural coefficients signs appear to be generally consistent with the expectations generated by the market model of rural agriculture. In particular the matrix of supply variable coefficient signs in the demand (wage) equations is supportive of the model, as wages appear to be sensitive to shifts in the supply of laborers for hire such that increases in the number of people participating in the agricultural labor market from any age-sex group reduce all agricultural wage rates, although not all coefficients are statistically significant. While the strongest supply impact on wages appears to come from shifts in female participation, the null hypothesis that an increase in female labor supply has equal negative effects on male and female wages cannot be rejected, as is consistent with but not implied by the model.

The supply equation structural estimates suggest that the relationship between the

[15] Because of severe multicollinearity, only own-wage rates were included in the supply equations. The wage-supply estimates thus represent the average relationships between sex-specific market labor supply and movements in the *set* of family wages.

[16] See Thomas Rothenberg and Charles Leenders. Because of the possibility that the *FIML* estimates will converge where the likelihood function is at a local rather than a global maximum, three-stage least squares was also employed. The parameter estimates obtained using this systems estimation method differed insignificantly from the *FIML* estimates.

TABLE 3—*FIML* STRUCTURAL COEFFICIENT ESTIMATES, INDIAN DISTRICTS, 1960–61

Independent Variable	Dependent Variable					
	W_M	W_W	W_C	λ_M	λ_W	λ_C
λ_M	−.055 (0.39)	−.0342 (0.24)	−.0501 (0.31)			
λ_W	−.0285 (2.49)	−.0321 (2.79)	−.0325 (2.30)			
λ_C	−.0295 (0.79)	−.0440 (1.30)	−.0395 (0.98)			
W_M				6.82 (0.96)		
W_W					18.94 (1.37)	
W_C						8.68 (1.07)
X_1	.0170 (4.69)	.0131 (3.65)	.0058 (1.39)	−.173 (1.24)	−.328 (1.53)	−.0441 (0.70)
X_2	.0058 (1.20)	.0031 (0.62)	−.0013 (0.23)	.368 (4.66)	.390 (3.47)	.0765 (2.09)
X_3	−.0064 (1.32)	−.0028 (0.54)	−.0050 (0.89)	.430 (4.63)	.606 (4.38)	.173 (3.30)
X_4	.0149 (4.51)			.153 (1.01)	.0092 (0.04)	−.136 (1.35)
X_5		.0122 (2.47)		−.217 (0.61)	−.474 (0.81)	−.0057 (0.03)
X_6	.0065 (0.65)			−.328 (0.89)	−.441 (0.80)	−.158 (1.02)
X_7		−.117 (0.26)		−3.54 (2.05)	−6.59 (2.78)	−2.58 (2.46)
X_8	−.0015 (0.26)	−.0004 (0.07)	−.0057 (0.83)	−.296 (1.92)	−.838 (2.82)	−.281 (1.79)
X_9	.0060 (3.41)	.0022 (1.23)	.0002 (0.10)			
X_{10}	.0003 (4.41)	.0002 (3.44)	.0001 (1.54)			
X_{11}	−.278 (2.21)	−.353 (2.88)	−.277 (1.95)			
X_{12}				−4.52 (1.09)	−8.52 (1.31)	−3.32 (1.80)
X_{13}				−.211 (0.84)	−.248 (0.65)	−.102 (0.97)
X_{14}				−.0382 (0.79)	−.284 (2.98)	−.0694 (1.39)
X_{15}				−12.49 (1.61)	−18.11 (1.65)	−4.74 (1.52)
X_{16}				.0411 (2.32)	−.0092 (0.33)	−.0229 (1.38)
Constant	1.64 (7.37)	1.16 (5.13)	0.92 (3.55)	−37.30 (2.50)	−49.74 (2.51)	−12.59 (2.01)
S.E.E.	.302	.302	.377	8.60	12.43	4.04

Note: Asymptotic t-values are shown in parentheses. Number of districts = 159.

quantity of laborers in each sex-age group supplying labor to the agricultural labor market and the level of wage rates is positive, although none of the wage coefficients are statistically significant by conventional standards. The coefficients of the land-distribution variables suggest, however, as expected, that decreases in landholding inequality and reductions in the proportion of landless households lower labor market

participation significantly. All of the coefficients of the nonfarm variables also display the expected negative signs, indicating that increases in demand factors in the nonagricultural sector draw laborers from farming (and raise rural wage rates), although all but two do not achieve statistical significance at the 10 percent level (one-tailed test).

While land size and the land-augmenting variables also have the expected positive effects on the demand for hired labor, the negative (but insignificant) signs displayed by the distributional coefficients, which indicate that where landholding inequality is greater the demand for hired labor (wages offered) is *lower* for all three groups, contradict the prediction of the purely competitive model under all scale economies assumptions. This latter result thus suggests that the negative relationships between landholding inequality and wage rates obtained in the reduced form may partly reflect the monopsonistic restrictions of wages and employment by relatively large landowners. Thus an equalization of the distribution of landholdings would appear to have a strong negative impact on the supply of agricultural wage labor as implied by the model, but negligible effects on hired labor demand with the net result that landholding inequality and rural wage rates are negatively and significantly associated in the reduced form.[17]

IV. Concluding Remarks

Prior theoretical models depicting rural household behavior in developing countries have employed differing restrictive assumptions with respect to labor-supply behavior and production which have led to contradictory predictions regarding the wage impact of land reforms. Most empirical studies of land reform, moreover, have not been based on a coherent theoretical framework, and have ignored wage effects.

In this paper I have investigated the wage effects of a redistribution of landholdings by formulating a general equilibrium, competitive market model embodying labor heterogeneity and more realistic labor-supply behavior. Although the model was constructed to be consistent with the important features of the agricultural labor market in India, it is sufficiently general so that it can be altered to suit structural conditions in the rural labor markets of other developing countries. The wage impact of a partial land reform was found to be theoretically indeterminate, due mainly to the assumption consistent with household-level India data that landowning labor exporting and importing households employ family labor, so that market labor-supply shifts are affected by opposing wealth-leisure effects. However, empirical results, which generally support the implications of the competitive model formulated, suggest that a redistribution of land from large- to small-farm households in India would raise agricultural wage levels significantly and thus benefit landless households, although sex differentials in rural wages would appear to widen.

[17]The derived reduced-form coefficients computed from the *FIML* structural parameters indicate that, given the actual distribution of landholdings in India in 1961–62, if a limit of fifty-one acres were placed on all farms and the "excess" distributed so that no landowning farm household would own less than 1.5 acres of arable land, wage rates would rise by 16.8, 17.0, and 14.1 percent for males, females, and children, respectively, in the absence of any other changes. See the author.

APPENDIX

Sources of Data:

Agricultural Wages in India, 1960-61: W_M, W_W, W_C.

Census of India, 1961: Part IIB—λ_M, λ_W, λ_C, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{16}. Part IIC—X_1, X_2, X_3, X_{15}. Part IVB—X_{12}, X_{13}, X_{14}.

Indian Agricultural Statistics, 1961-62 and 1962-63: X_9, X_{11}.

REFERENCES

K. J. Arrow, H. Block, and L. Hurwicz, "The Stability of the Competitive Equilibrium, Part II," *Econometrica*, Sept. 1959, *27*, 82–109.

P. K. Bardhan, "Size, Productivity, and Returns to Scale: An Analysis of Farm Level Data in Indian Agriculture," *J. Polit. Econ.*, Nov./Dec. 1973, *81*, 1370-386.

——— **and T. N. Srinivasan,** "Cropsharing Tenancy in Agriculture: A Theoretical and Empirical Analysis," *Amer. Econ. Rev.*, Mar. 1971, *61*, 48-64.

C. Bell and P. Zusman, "A Bargaining Theoretic Approach to Cropsharing Contracts," *Amer. Econ. Rev.*, Sept. 1976, *66*, 578-88.

R. A. Berry, "Land Reform and the Agricultural Income Distribution," *Pakistan Develop. Rev.*, Spring 1971, *11*, 30-44.

Ester Boserup, *Women's Role in Economic Development*, London 1970.

M. Gersovitz, "Land Reform: Some Theoretical Considerations," *J. Develop. Stud.*, Oct. 1976, *12*, 79-81.

W. A. Lewis, "Economic Development with Unlimited Supplies of Labor," *Manchester Sch. Econ. Stud.*, May 1954, *22*, 139-91.

R. Mabro, "Employment and Wages in Dual Agriculture," *Oxford Econ. Pap.*, Nov. 1971, *23*, 401-17.

D. Mazumdar, "The Theory of Share-Cropping with Labour Market Dualism," *Economica*, Aug. 1975, *42*, 261-71.

D. Newbery, "Cropsharing Tenancy in Agriculture: Comment," *Amer. Econ. Rev.*, Dec. 1974, *64*, 1060-66.

M. A. Rahman, "Farm Size, Efficiency and the Socioeconomics of Land Distribution," *Bangladesh Econ. Stud.*, July 1975, *3*, 301-18.

G. Ranis and J. Fei, "A Theory of Economic Development," *Amer. Econ. Rev.*, Sept. 1961, *51*, 533-65.

G. B. Rodgers, "Nutritionally Based Wage Determination in the Low-Income Labour Market," *Oxford Econ. Pap.*, Mar. 1975, *27*, 61-81.

M. R. Rosenzweig, "Rural Wages, Labor Supply and Land Reform: A Theoretical and Empirical Analysis," disc. paper no. 70, Res. Program Develop. Stud., Princeton Univ., Dec. 1977.

——— **and R. Evenson,** "Fertility, Schooling and the Economic Contribution of Children in Rural India: An Econometric Analysis," *Econometrica*, July 1977, *45*, 1065 80.

T. Rothenberg and C. Leenders, "Efficient Estimation of Simultaneous Equations Systems," *Econometrica*, Jan./Apr. 1964, *32*, 57-76.

A. K. Sen, "Peasants and Dualism With or Without Surplus Labor," *J. Polit. Econ.*, Oct. 1966, *74*, 425-50.

S. Wellisz, "Resource Allocation in Traditional Agriculture: A Study of Andhra Pradesh," *J. Polit. Econ.*, July/Aug. 1970, *78*, 655-84.

Directorate of Economics and Statistics, *Agricultural Wages in India, 1960-61*, New Delhi 1964.

———, *Indian Agricultural Statistics, 1961-62 and 1962-63*, Vol. II, New Delhi 1970.

Office of the Registrar General, *Census of India, 1961*, New Delhi 1965.

Time in School: The Case of the Prudent Patron

By THOMAS JOHNSON*

Given my capabilities and opportunities, how much time should I spend in school? That is a question which confronts almost every person in modern societies at some time in their lives; and the effort to provide the answer to individuals occupies counselors, testers, and absorbs much of the effort of teachers. Economists also try to turn the question around and make inferences about personal characteristics, and predict responses to changes in policies or market conditions. From whichever side we view the question we soon realize that, as stated, the question is not precise enough for our purpose.

How do you define and measure capabilities and opportunities? By "time," do you mean years spent "in school," or the fraction of each day spent in class and studying? The growing human capital tradition in economics abstracts heroically and embodies capabilities in a few parameters of some production function and an initial stock of human capital. Opportunities are represented by wages or rental rates, prices of purchased commodities, and deterioration and/or growth functions which are external to the individual. Similarly, "time in school" studies have focused almost exclusively upon the years spent. In this paper I try to expand the scope of the model to include analysis of the decision of a "full-time" student to work in the labor market. Increasing the number of variables in the model and exploring different institutional assumptions give results which allow us to estimate empirically the relative importance of own-human capital and purchased inputs to the human capital production function of students. One estimation technique yields estimates of the average importance for all students, while an alternative technique yields estimates for individual students who work in the market. However, it is well to retrace briefly the path of those who have brought us this far before exploring new ground.

Almost by definition studies in human capital have from the beginning been concerned with optimal accumulation. In particular, Gary Becker (1964, 1967) demonstrated the power of neoclassical notions of economizing choice in applications to questions of the acquisition and employment of skills. Yoram Ben-Porath initiated a line of inquiry into the use of optimal control theory in the analysis of the accumulation of human capital over the individual's lifetime. This original paper assumed perfect markets for borrowing, lending, and renting human capital by an individual restricted to dividing his time between earnings and investment in human capital without any direct gifts, allowances, or stipends. The literature now includes papers that explore Ben-Porath's original problem in greater depth (see William Haley, 1973; Lee Lillard, 1973, 1974), including opportunities for choice of leisure (see Harl Ryder, Frank Stafford, and Paula Stephan), eliminating the opportunities for loans (see T. D. Wallace and Loren Ihnen), and including gifts to support specialization (see Haley, 1976).

The first departure of the present analysis from previous work is to assume that the student specializing in the production of human capital can earn only a relatively low wage which does not increase while he is in school. I maintain the assumption that the rental rate on human capital is constant after the period of specialization. The assumption of a constant (and low) wage rate while specializing amounts to assuming that the student cannot market

*Professor of economics and statistics, North Carolina State University at Raleigh. This is paper no. 4594 of the journal series of the North Carolina Agricultural Experiment Station, Raleigh. I wish especially to thank Loren Ihnen and the managing editor for suggestions for clarification of the presentation of these results.

the services of his entire stock of human capital.

The second departure is to include receipt of an allowance, which is a flow at a constant rate while the student specializes. Specialization is now defined to mean that the total of earnings and the allowance just equals the value of purchased inputs to the production of human capital, where purchased inputs can be considered to include a "subsistence" level of living. These somewhat unusual assumptions are discussed more fully below.

A unique result of these assumptions is a relation for the determination of whether or not a full-time student participates in the labor market. The model student will not participate in the labor market if his annual allowance is greater than β_2/β_1 times the annual potential (full-time) wage rate he could earn while a full-time student. The term β_1 is the coefficient of own-human capital, and β_2 is the coefficient of purchased inputs in a Cobb-Douglas production function for human capital. Thus, by analysis of the dichotomous decision whether or not to work for wages while in school, we can now obtain an estimate of the β_2/β_1 ratio. Previously, only estimates of $\beta_1 + \beta_2$ have been possible.

Logit analysis is applied to the work decision data of 105 persons who graduated from high school in 1971 and attended public four-year colleges and universities in 1971-72. This analysis indicates that loans have an impact of approximately the same magnitude and statistical significance as direct allowances. Estimates of the ratio β_1/β_2 are between 5 and 7, emphasizing the importance of own-human capital relative to purchased inputs in producing more human capital in school.

A numerical sensitivity analysis has been performed in which changes in the allowance, wage, ability, and β_1/β_2 ratio are related to the age at which full-time schooling ends. The results imply that, even when allowances and the in-school wage rate would have equal effect on the "work during school" decision, the decision to leave full-time schooling is much more sensitive to an increase in the allowance than to an increase in the in-school wage rate. However, increasing the allowance, other things equal, has opposite effects on earnings early in life and late in life. Increasing the allowance to induce continued schooling causes the lifetime income profile to become more nearly constant.

I. The Model

The services from the stock of human capital a person possesses are divided between the production of more human capital and the production of income. At each moment the fraction of effort spent investing is labeled s_i, while the fraction spent earning is labeled s_m with $s_i + s_m = 1$.[1] To produce human capital, the person combines his own human capital, $K_i(t) = s_i(t)E(t)$, and purchased inputs, $D(t)$, in a production function[2]

$$Q(t) = \beta_0 K_i^{\beta_1} D^{\beta_2} \tag{1}$$

The model results will show that it is optimal for a person to continue producing human capital until the end of life so that K_i and D are positive throughout life (except at the last instant).

The model decision maker starts life at $t = 0$ which corresponds to the time at which he plans his own life. There are two possible phases for the life plan. In phase I, the individual may for a time specialize in the production of human capital, or he may begin market work immediately with no period of specialization beyond $t = 0$. We will say that the individual is specializing in the production of human capital when all current income, $Y(t)$, is spent on pur-

[1]The advantages of this specification are that it is most useful for extending the analysis to multiple activities and activities which do not depend upon the level of the stock, and it more clearly shows the connection with Gary Becker's discussion of the allocation of time (1975). Symbols are defined in Table 1.

[2]A technical appendix made available to the reader on request contains the derivation of solutions to the model. Before I can solve for the entire life cycle results, the assumptions made imply the Cobb-Douglas form for Q.

TABLE 1—LIST OF VARIABLES AND PARAMETERS

Variable	Description
Endogenous Variables	
$Y(t)$	actual income
$E(t)$	human capital stock
$s_i(t)$	the fraction of the effort at t spent investing
$s_m(t) = 1 - s_i(t)$	the fraction of the effort at t spent in the labor market
$K_i(t) = s_i(t)E(t)$	human capital invested
$D(t)$	purchased educational inputs
$Q(t)$	produced human capital
$\lambda(t)$	shadow price of net additions to the human capital stock at t
t_g	time at which specialization ends
Exogenous Variables and Parameters	
w	wage rate while specializing
A	allowance while specializing
A	allowance while specializing
R	rental rate on human capital after specialization
δ	rate of deterioration of $E(t)$
P	price of purchased inputs
$\beta_0, \beta_1, \beta_2$	production function parameters
E_0	initial stock of human capital
T	length of planning horizon
r	rate of time discount
t	the number of years since the individual decided upon a life cycle path of learning and earning
Working Constants	
$\Delta = 1 - \beta_1 - \beta_2$	
$q_I = Q(s_i^I, D^I)$	
$q_{II} = Q(1, c_{II})$	
$c_I = \begin{cases} \beta_2 w/\beta_1 P & \text{when } s_i^I < 1 \\ A/P & \text{when } s_i^I = 1 \end{cases}$	
$c_{II} = R\beta_2/P\beta_1$	
$q_{II} = \beta_0\beta_1(R\beta_2/P\beta_1)^{\beta_2}$	

chased inputs D. No loans are available for the purchase of D, and the person may specialize while working part time to finance purchases of D.[3] If the individual ever specializes it will be at the beginning of the life plan and, therefore, the specialization phase is contained in phase I. In the second phase, phase II, the individual does not spend all income on D.

During phase I, the available wage rate is w which is independent of the stock of human capital; but an allowance A is also received. The allowance is an unearned flow of funds at a constant rate during phase I. In phase II the available wage rate is $RE(t)$, proportional to the stock of human capital.[4] Actual income at time t may be written

$$(2) \quad Y(t) = \begin{cases} ws_m(t) + A & \text{Phase I} \\ RE(t)s_m(t) & \text{Phase II} \end{cases}$$

In the "perfect capital market" models of Ben-Porath and Haley (1973) it is clear that $s_m = 0$ during specialization. Conversely, in the "no-investment loans" model of Wallace and Ihnen $s_m > 0$ for all t, since purchased inputs are required to produce human capital and earnings are the only source of income. In the model of this paper, with the allowance and the requirement that inputs be purchased out of current income, s_m may be greater than or equal to zero during phase I. However, a result of the model is that s_m will be constant during phase I. In this model there are two possible cases during phase I. Phase I, case 1 corresponds to $s_m > 0$ ($s_i < 1$, the student works in the market). Similarly, phase I, case 2 corresponds to $s_m = 0$ ($s_i = 1$, the student does not work in the market).

The price of D is a constant P, and the end of phase I, time of graduation t_g, may be defined by

$$(3) \quad \begin{aligned} Y(t) - PD(t) &= 0 \qquad \text{Phase I, } t \leq t_g \\ Y(t) - PD(t) &> 0 \qquad \text{Phase II, } t > t_g \end{aligned}$$

Figure 1 presents a comparison of income $Y(t)$ and expenditures $PD(t)$ over a life cycle. In this illustration A = \$3,000 per year and earnings during phase I are \$750 per year with a wage rate w = \$4,500 per year (\$2.25 per hour for a 2,000-hour year).

[3]The purchased inputs include a subsistence level of consumption. To keep the separation theorem, making income maximization equivalent to utility maximization, we follow Wallace and Ihnen in our "no investment loans" assumption. Loans are not available to finance investment but are available to finance consumption (see Wallace and Ihnen, p. 139) beyond the subsistence level.

[4]Thus, it is clear that this wage corresponds to the "potential income" of other studies. Both refer to the amount which could be earned if the individual worked one unit of time.

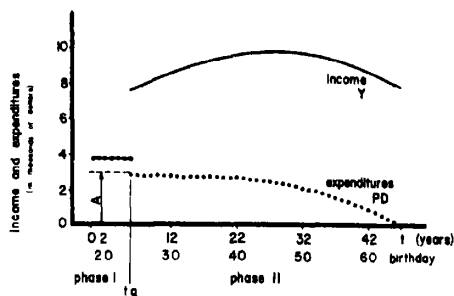



FIGURE 1 AN EXAMPLE OF THE FLOWS OF INCOME AND EXPENDITURES

The assumption of a constant w is supported by casual empiricism; while the college senior has two times or more the human capital of a college freshman,[5] the wage rate for waiting on tables (or similar student jobs) increases very little. The case is easily rationalized by noting that student job markets are often geographically isolated and are of insufficient size to support extensive specialization, and that the managerial costs of organizing part-time workers reduces the marginal product. Furthermore, as shown by the author (1975), employers may view schooling investments in human capital which is not firm specific to be a form of moonlighting. More formally, a constant w is equivalent to assuming that a full-time student can rent only some constant fraction of the human capital with which he *starts* college.

The rationalization for the constancy of A and the condition under which it is given also are not derived from an explicit maximization model. To obtain A as an explicit result would require a complete model of intergenerational transfer within the family and of the grants economy for scholarships. Instead, as with the wage rate assumption, more casual support is relied upon. The scholarships with which we are familiar increase very little as the recipient advances through his college career. Perhaps this is connected with the practice of educational institutions of charging the same price per course without regard to the level of the course or the human capital stock of the student.[6]

The prudence of the patron is evidenced by the requirement that all income be spent on schooling as the condition for receipt of the allowance. The patron does not require the student to work in the low-wage market; the low wage is caused by the separation of markets. But why would the patron impose the expenditure restriction on the student? Why impose any restrictions, that is, why not just make an unrestricted transfer? An answer to the last question may be that the benefactor believes that his judgment of the welfare of the recipient is superior and believes that the beneficiary would not follow this judgment unless constrained to do so. However, the patron will find it difficult to determine when his grant is being used in accordance with his wishes. It will be especially difficult to determine the amount of the beneficiaries' own expenditures which have been displaced from the project the benefactor wishes to support. Therefore, as with other grant programs, we expect to find some rather arbitrary rules imposed as conditions for the grants. The conditions imposed by the "prudent patron" are suggested as a reasonable abstraction of the

[5]These proportions are implied by the estimates by the author (1970), Table 1, pp. 556–57. For example, the base human capital B of a white male in the North entering college is \$3,920 while the base human capital B of a similar person entering graduate school is \$10,770. Preliminary analysis of the data for the high school class of 1972 (see Research Triangle Institute, 1975b), the follow-up to the data used in this paper, shows that while nonstudent's wages increased by a real 8.9 percent from October 1972 to October 1973, the wage of full-time students in four-year colleges and universities increased by a real 2.3 percent.

[6]The change in measured allowance for the high school class of 1972 between their first and second years in college can be calculated from the data of the second follow-up of the class of 1972. Results from 1,259 persons from the class of 1972 who attended the same college for their first two years show a mean reduction of approximately 20 percent (in 1972 dollars) in allowance between the first and second year in college. Therefore, these data do not support the assumption of a constant allowance, although the discrepancy is in the opposite direction from Haley's assumption of allowance proportional to the quantity of human capital produced. However, for this study the assumption of constant allowance is retained.

conditions imposed by actual benefactors and grant administrators in a world of imperfect information.[7]

Assuming that human capital deteriorates at a constant rate, $0 < \delta < 1$,

$$\frac{dE(t)}{dt} = Q(t) - \delta E(t) \tag{4}$$

gives the rate of net addition to the stock of human capital.

Assume that the individual has a constant rate of time discount r, and that his objective is to maximize the present value of his actual income less the cost of purchased inputs to the production of human capital. This maximization is taken over a fixed planning horizon T, with no utility assigned to stocks accumulated at T. Thus, T represents the moment of death. Since no utility of leisure or disutility of work is introduced into the objective function, and the shadow price of E goes to zero at T, the person will approach complete devotion to the labor market, $s_m = 1$, as t approaches T. Therefore, the problem is:

$$\text{(5a)} \quad \max J = \int_0^T [Y(t) - PD(t)]e^{-rt}dt$$

with respect to $\overrightarrow{s_m(t)}$, $\overrightarrow{D(t)}$, and subject to:

$$\text{(5b)} \quad dE/dt = Q(t) - \delta E(t)$$

$$\text{(5c)} \quad s_i(t) + s_m(t) = 1$$

$$\text{(5d)} \quad Y(t) - PD(t) \geq 0,$$
$$s_i(t) \geq 0, s_m(t) \geq 0, D(t) \geq 0$$

with $Q(t)$ defined in (1) and $Y(t)$ defined in (2).

This is an optimal control problem which is best solved by application of the maximum principle. The Hamiltonian corresponding to problem (5) is

$$\text{(6)} \quad H(s_i, s_m, D, \lambda; E) = [Y(t) - PD(t)]e^{-rt} + \lambda(t)[Q(t) - \delta E(t)]$$

The problem equivalent to (5) is then, for each t,[8]

$$\text{(7a)} \quad \max H(s_i, s_m, D, \lambda; E)$$

with respect to s_m, D, and subject to:

$$\text{(7b)} \quad \partial H/\partial E = -d\lambda/dt$$

$$\text{(7c)} \quad \partial H/\partial \lambda = dE/dt$$

$$\text{(7d)} \quad s_i + s_m - 1 = 0$$

$$\text{(7e)} \quad PD - Y \leq 0$$

$$\text{(7f)} \quad E(T)\lambda(T) = 0, s_i \geq 0, s_m \geq 0, D \geq 0$$

A. *Phase I*

While the student specializes in the production of human capital, the following conditions hold by definition

$$\text{(8a)} \quad 0 \leq t \leq t_g$$

$$\text{(8b)} \quad Y = ws_m + A$$

$$\text{(8c)} \quad Y = PD$$

$$\text{(8d)} \quad \partial Y/\partial E = 0$$

$$\text{(8e)} \quad \partial Y/\partial s_m = w$$

$$\text{(8f)} \quad s_m \geq 0$$

During phase I all income comes from the allowance A plus earnings at the constant wage rate w and must be spent on purchased inputs to human capital production. Since all effort must be divided between investment and market work, (5c), it follows immediately from (8b) and (8c) that the optimum quantity of purchased inputs is a linear function of s_i

$$D^I = \frac{w + A}{P} - \frac{w}{P} s_i^I \tag{9}$$

for both cases.

[7]Of course we must recognize the arbitrariness of the definition of subsistence to be included in the allowed cost. Some scholarships stipulate that the student must have no employment outside the college or university. However, if some administrator approves, the scholar may have earnings within limits. Some patrons may prohibit market work, require $s_m = 0$. However, the assumption made in the text yields more analytical power, and I believe it is fairly realistic.

[8]See Michael Intriligator, pp. 344–53, or George Hadley and Murray C. Kemp, pp. 238–49, for useful discussions. The explicit recording of the dependent variable t is now dropped since the equations are becoming more complicated. The reader is referred to Table 1 if there is any question about what is a function of t. The details of the solution to problem (7) are available upon request.

The necessary conditions for the solution of problem (7) imply

$$(10)\quad s_l^I = \begin{cases} \dfrac{w + A}{w + w\beta_2/\beta_1} & \text{if } \beta_1 A < \beta_2 w \\ 1 & \text{if } \beta_1 A \geq \beta_2 w \end{cases}$$

For those students who work in the market the ratio of elasticities is given by

$$(11)\quad \frac{\beta_1}{\beta_2} = \frac{ws_l^I}{A + Ws_m^I} = \frac{\text{foregone earnings at wage } w}{\text{total income}}$$

Both (10) and (11) provide opportunities to estimate the ratio β_1/β_2. Estimates may be obtained from (10) for a population by logit analysis of the dichotomous choice of individual students to participate or not to participate in the market. Estimates may be obtained from (11) for individual students who do chose to participate in the market.

During phase I the stock of human capital E^I and its shadow price λ^I are given by

(12)

$$E^I = [\frac{q_I}{\delta} + (E_o^{1-\beta_1} - \frac{q_I}{\delta})e^{-\delta(1-\beta_1)t}]^{1/(1-\beta_1)}$$

and

$$(13)\quad \lambda^I = \lambda_g(E_g/E^I)^{\beta_1}e^{-\delta(1-\beta_1)(t_g-t)}$$

The quantities E_g and λ_g[9] denote the values at t_g, the age at which phase I ends, and E_o denotes the stock of human capital at $t = 0$.

B. *Phase II*

When the student quits specialization he has to give up his allowance, but he can then rent all of his human capital for the rate R. His wage rate is then RE instead of being restricted to w. Thus the following conditions define phase II:

$$(14a)\quad t_g < t \leq T$$

$$(14b)\quad Y = s_m RE$$

$$(14c)\quad Y > PD$$

$$(14d)\quad \partial Y/\partial E = s_m R$$

$$(14e)\quad \partial Y/\partial s_m = RE$$

$$(14f)\quad s_m > 0$$

During phase II it is more convenient to write the solution[10] in terms of $K_l^{II} = s_l^{II}E^{II}$:

$$(15)\quad K_l^{II} = [\tilde{q}_{II}\frac{(1 - e^{-(r+\delta)(T-t)})}{r + \delta}]^{1/\Delta}$$

$$E^{II} = \frac{R\beta_2}{P\beta_1}K_l^{II}$$

$$E^{II} = E_g e^{-\delta(t-t_g)} + q_{II}[\frac{1}{r + \delta}]^{1/\Delta}(\tilde{q}_{II})^{(1-\Delta)/\Delta}e^{\delta(T-t)}$$
$$[B(v;\frac{\delta}{r + \delta}, \frac{1}{\Delta}) - B(v_g; \frac{\delta}{r + \delta}, \frac{1}{\Delta})]$$

$$\lambda^{II} = \frac{R}{r + \delta}e^{-rt}(1 - e^{-(r+\delta)(T-t)})$$

Making comparisons between phases I and II, there is a striking contrast between the complexity of the expressions for E^I and E^{II}, and between λ^I and λ^{II}. The reason for this is intuitively appealing. Since E is "tied" to E_o at $t = 0$ at the beginning of phase I the identifying restriction on E must be carried forward through the switching point t_g to identify E during phase II. The stock of human capital with which the maximizer starts is known; and while at any moment he could change phases he would have to start that new phase with the stock he has accumulated up to that moment.

Conversely, λ is tied to zero at $t = T$ the end of phase II: "You can't take it with you." This identifying restriction on λ must be carried back through the switching point t_g to identify λ during phase I. At each moment the value of an addition to the stock

[9]The boundary conditions are represented by λ_g and E_g. To obtain the solution for λ we must work back from $t = T$, since the final value is $\lambda(T) = 0$.

[10]The *beta* function in (15) is $B(\alpha,\beta) = \int_o^1 u^{\alpha-1}(1 - u)^{\beta-1}\,du$ and the incomplete *beta* function is $B(v,\alpha,\beta) = \int_o^v u^{\alpha-1}(1 - u)^{\beta-1}\,du$ for $0 \leq v \leq 1$. Karl Pearson tabulated the ratio $B(v;\alpha,\beta)/B(\alpha,\beta)$. In computations of E^{II} a series approximation to $B(v;\alpha,\beta)$ is employed.

of human capital depends upon the use that will be made of it in the future.

C. *The End of the Phase of Specialization*

The maximum principle tells us that the maximizer can achieve his objective of equation (5) if at each moment he makes the choice to maximize H subject to the constraints shown in (7). The solutions satisfy the restrictions so the optimal criterion for switching phases is the moment the Hamiltonian for phase II exceeds the Hamiltonian for phase I. To determine the optimal switching point, only the comparison of $H^I(t)$ and $H^{II}(t)$ is needed (both are computed under the assumption the switch is made at t).

Equating $H^I(t_g)$ and $H^{II}(t_g)$, and substituting for Y, s_i, D, E, and λ into equation (6), yields an implicit function in t_g. Since there is no discontinuity in E or λ, $E(t_g)$ is obtained from the phase I solution while $\lambda(t_g)$ is obtained from the phase II solution.[11] However, Y, s_i, and D are different for H^I and H^{II}. Thus the implicit function for t_g is

$$H^I(t_g) - H^{II}(t_g) = 0 \tag{16}$$

Applying the implicit function differentiation theorem, it can be shown that for long work lives, t_g varies inversely with E_o. Similarly, it can be shown that t_g varies inversely with R if $\beta_1 \leq (r + \delta)E_g/Q'_g$, i.e., persons with relatively low coefficients on own-human capital will leave school earlier as the postschool rental rate increases.[12]

Also, it can be shown that when the student is in the labor market the relative responsiveness of t_g to w and A is

$$\frac{\partial t_g/\partial w}{\partial t_g/\partial A} = 1 - \frac{\beta_1}{1 - \Delta}\left(1 + \frac{A}{w}\right) \tag{17}$$

which is always less than one. The response of t_g to w would be negative if w were less than $\beta_1 A/\beta_2$, and the student worked; but then the student would not work. Thus the response of t_g to the in-school wage rate w is nonnegative but less than the responsiveness to A.

A series of numerical calculations of the optimal t_g was performed to investigate the sensitivity of t_g and income patterns to changes in the parameters and the initial conditions. The results are tabulated in an appendix available from the author upon request. These results indicate that increasing the Hicks-neutral ability parameter β_o has very little effect on t_g; changes in t_g resulting from changing β_o from 40 to 50 are within the range of computational imprecision. Increasing ability β_0 increases earnings later in life, but has little effect before birthday 25.[13]

Increasing the allowance A, other parameters held constant, has *opposite* effects on earnings early and late in life. The higher the allowance the greater the earnings are between leaving school and birthday 35; but, earnings after 45 *decrease* with an increase in the allowance. Inducing a person of given ability to remain in school longer "twists" his lifetime earnings profile.

Increasing β_2 with $\beta_1 + \beta_2 = 0.5$, i.e., increasing the relative productivity of purchased inputs, unambiguously reduces the years spent in school and the proportion of effort devoted to schooling while specializing.

II. Estimation of β_1 and β_2

Even after assuming that the price of purchased inputs is a constant throughout life

[11]The expressions are so complicated that I have been unsuccessful in attempts to determine analytically under what conditions $dH^I_g/dt < dH^{II}_g/dt$. However, numerical calculations with a Cobb-Douglas production function Q indicate that the Hamiltonians are well behaved in the empirically interesting range. It might be more realistic to require that t_g come only at the ends of semesters or academic years. However, that restriction would carry us into discrete optimal control problems for which general analytical results are more difficult to obtain. The numerical calculations obtain t_g to the nearest hundreth of a year.

[12]A necessary condition is too complicated to be of interest although for large values of $T - t_g$ it is not necessary for β_1 to be "much" greater than $(r + \delta)E_g/Q'_g$ to cause $\partial t_g/\partial R > 0$.

[13]Here birthday is used to represent age in the popular sense. In the calculations $t = 0$ was taken to be at birthday 18.

and that the parameters of the production function for human capital are also the same at each age, we still face the difficulties, practical and conceptual, of observing the quantity D and price P of purchased inputs. Haley has responded to these difficulties by dropping purchased inputs from his production function. However, if the aggregate data Haley uses represent an optimizing individual, his estimate of the production function parameter for s_t or $E(t)$ is an unbiased estimate of $\beta_1 + \beta_2$.[14]

My criterion for entering the labor market while specializing offers promise of a second equation in β_1 and β_2. Equation (11) says that a student should work if and only if $\beta_1 A < \beta_2 w$. We can, by reasonable questionnaires, find out if a student works any during the year or not, how much nonwage income he receives, and what he perceives to be his wage rate. Thus we might try to explain the dichotomous variable, work in the market, by the observed allowance (including state subsidies) and the potential wage rate while a student. Of course in the real world we don't expect that choices are cut with that sharp a razor; tastes and information differences will cause variation.

Therefore the random utility model and conditional logit analysis is employed to estimate the response of labor force participation to changes in allowance and wages. Following Daniel McFadden (1973), an individual has a utility function

$$U = V(a,c) + e(a,x) \tag{18}$$

where V is nonstochastic and is "representative" of the population, and e is the stochastic reflection of the idiosyncracies of the individual's tastes for the available alternative with attribute x; and the individual has measured attributes a. Assuming that V is linear in both the parameters and the variables x and a we would estimate the parameters of

$$V(a,x) = \gamma_o + \gamma_1 x_1 + \ldots + \gamma_m x_m + \gamma_{m+1} a_1 + \ldots + \gamma_k a_{k-m} \tag{19}$$

from the observations on individuals.[15]

In our problem the x's are measures of job and school characteristics; $a_1 = A$ and $a_2 = w$, with the other a's including such potential shifters as student loans, state support of the school, work experience, marital status, scholastic aptitude, family income, etc. The formal model holds P and w constant. These additional variables are included in an informal attempt to hold relative prices constant between persons in the sample. Keep in mind that our model predicts that scholastic aptitude (as measured by β_o and/or E_o) should not affect the labor force participation decision in phase I. The estimates of the coefficients of A and w in equation (19) are not direct estimates of β_1 and β_2.[16] However the ratio of the γ's give estimates of the relative response to A and w so that we can obtain an estimate of β_1/β_2.

The data employed are from the "Preliminary National Longitudinal Study: Class of 1971" (Research Triangle Institute, 1975a). Of the 523 sampled persons from the high school graduating class of 1971, there are 105 usable returns for those who attended public four-year institutions in the

[14]Suppose that the individual follows the criteria of equation (15). During phase II

$$Q = \beta_o(\beta_2 R/\beta_1 P)^{\beta_2}(s_t E)^{\beta_1+\beta_2}$$

and when Haley estimates the parameters of

$$Q = b_o(s_t E)^b$$

his estimates of b are unbiased estimates of $\beta_1 + \beta_2$ as surely as they are unbiased estimates of b

[15]As in ordinary least squares regression a crucial assumption is that the parameters are constants across the population; any changes are to be specified by some variable included in the model. Also, in logit regression the error terms are assumed to be of the same form (constant parameters) between populations. Because of the power of our theory in giving two ways to estimate β_1/β_2 we have evidence that these two parameters are not constant across the entire population. The econometric problems thus presented are a subject for future research.

[16]The magnitude of the γ's will depend upon the rate at which students respond to the criterion; if response were "perfect" according to the model, then participants and nonparticipants would be completely separated by the criterion and the estimation procedure would not yield finite estimates of the γ's.

fall of 1971. The variables used are defined in the footnote in Table 2. The indirect definition of market work is required because responses to the direct question on work indicates that many respondents misinterpreted the year to which the question referred.

One surprising result is that state aid to the institution shows an effect in the opposite direction to direct aid to the student. It was thought that this result might have been a result of schools in larger urban centers receiving more state support per student. In larger cities, or cities with a smaller

TABLE 2—SUMMARY OF LOGIT RESULTS

	Regression Number							
	1	2	3	4	5	6	7	8
Restriction	$\alpha_9 = 0$	$\alpha_6 = \dots = \alpha_9 = 0$	$\alpha_5 = \dots = \alpha_9 = 0$	$\alpha_5 = \dots = \alpha_9 = 0$ $\alpha_1 = \alpha_2$	α $\alpha_9 = 0$	$\alpha_2 = \alpha_4 = \dots = \alpha_9 = 0$	$\alpha_5 = 0$	$\alpha_5 = \dots = \alpha_8 = 0$
Constant	0.2859	0.5391	0.4209	0.4196	0.6874	0.6137	0 8618	0.4007
(α_0)	(0 6678)[a]	(0.5100)	(0.4975)	(0.4995)	(0.4255)	(0.4173)	(0.6739)	(0 4983)
Allowance	−0.3342	−0.3586	−0.3777	−0.3846	−0.3332	−0.2845	−0.3588	−0.3754
(α_1)	(0.1449)	(0.1388)	(0.1379)	(0.1362)	(0 1298)	(0 1214)	(0 1456)	(0 1386)
Loans	−0.4102	−0.4708	−0.4957	−0 3846	−0 4036		−0.4005	−0.4834
(α_2)	(0 3120)	(0.3001)	(0.2974)	(0.1362)	(0 2877)		(0 3128)	(0.2999)
Wage	0.05217	0.05696	0 05578	0 05394	0.07236	0.05085	0.05762	0.05873
(α_3)	(0 1015)	(0.09533)	(0.09506)	(0.09533)	(0.08912)	(0.08769)	(0.1006)	(0.09521)
State Aid	0 2093	0.2494	0 2753	0 2702			0.2052	0 2568
(α_4)	(0.1796)	(0.1825)	(0 1810)	(0.1810)			(0.1855)	(0.1886)
S/C	−0 3759	−0.3200						
(α_5)	(0.3423)	(0.3177)						
Sex	0 3506						0 3121	
(α_6)	(0.2463)						(0.2402)	
Race	−0.1469						−0.07562	
(α_7)	(0.4586)						(0 4578)	
Marriage	0.1986						0 3101	
(α_8)	(0 3948)						(0.3858)	
City Size							0.5610	0.3305
(α_9)							(1.028)	(0 9686)
Log Likelihood	−56.445	−57.685	−58.196	−58.285	−59.754	−60.741	−56 907	−58.136
χ^2		2.48	3 50	0.18	3.116[b]	5.09[b]		2.46
df		3	4	1	1	2		3
Comparison[c]		2 to 1	3 to 1	4 to 3	5 to 3	6 to 3		8 to 7
$\beta_1/\beta_2 = \alpha_1/\alpha_3$	6.41	6.30	6.77	7.13	4.60	5.59	6.23	6.39

Note: To calculate probabilities, multiply coefficients by 2

[a] Asymptotic standard error

[b] Significant at the 10 percent level.

[c] Identification of the regressions compared, for example, 2 to 1 means that regression number 2 is compared to regression number 1.

Definition of Variables:

Dependent variable (work) = 1 if the student reported financing of the first year of college from earnings while in school, summer earnings, or savings and = 0 otherwise.

A (allowance) = payments directly to the student of scholarships, grants, aid from other persons, and other aid, measured in thousands of dollars per year.

Loans = sum of all classes of loans reported in thousands of dollars per year

w (wage) = (a) (2,000 × hourly wage)/1,000 if the student worked during October 1972 or if the most recent wage is reported. (b) (2,000 × estimated hourly wage)/1,000 where the wage is estimated by regressing the hourly wage or the sex, race, work in high school, and type of school attended by those reporting wage rates.

State aid = the total state appropriation to the school attended divided by the number of students reported for that institution in thousands of dollars per year.

S/C = student population divided by the population of the city in which the school is located.

City size = the population of the city in which the school is located measured in thousands.

Sex = 0 if female, 1 if male.

Race = 0 if other, 1 if white.

Marriage = 0 if married, 1 if single or separated.

student/population ratio (S/C), one would expect less effect of nonprice job rationing. The combination of these two effects associated with city size might cause a spurious negative correlation between state aid and student labor force participation. Of the two variables tried, the S/C variable entered with a statistically more significant coefficient than city size (compare regressions 1 and 7). However, neither variable significantly affected the magnitude or significance of the state aid variable. Thus we are left with a puzzle on the effect of state aid to the school on the labor force participation of the students.

Allowance performs as expected. Increasing the allowance[17] decreases the probability of labor force participation and the effect is statistically significant. Surprisingly, loans have practically the same effect as grants. Regression 4 shows that there is no significant loss in the likelihood function from combining grants and loans; apparently the returns to schooling are sufficiently greater than the interest rate charged to make loans nearly equivalent to grants. Although the asymptotic standard error of the estimate of the wage coefficient is large, the estimated ratio β_1/β_2 is quite robust to changing specifications.

Equation (11) provides an alternative procedure to estimate β_1/β_2 for each individual who works in the market while a student. The same data set provides observations on 42 students for which equation (11) could be employed. For these 42 students the mean value of β_1/β_2 is 2.57 with a standard error of 0.37 and the range is -0.134 to 12.714. Deleting the high and low outliers, the mean value of β_1/β_2 for 40 students is 2.38 with a standard error of 0.28 and the range is 0.311 to 7.71. Therefore these results indicate that working students have a lower than average coefficient on own-human capital relative to the coefficient on purchased inputs. Furthermore, the relative importance of own-human capital and purchased inputs seems to be highly variable between persons.[18]

III. Summary

This paper explores the properties of a life cycle model of human capital accumulation under the assumptions that the individual cannot borrow to finance his schooling, but may receive an allowance while specializing. This allowance is constant, regardless of the amount of human capital accumulated, and should be interpreted to include subsidies to tuition as well as funds to cover a "subsistence" level of living. It is also assumed, realistically I think, that the specializing student cannot rent his human capital for the same rate as he could if he were not specializing. This produces the empirically familiar "notch" in the fraction of effort spent investing in human capital.

Numerical estimates of the responsiveness to changes in parameter values indicated that, in most cases, increasing the student wage had little effect on both time spent in school and income later in life. On the other hand, increasing the allowance did increase the time spent in school. However, with ability and human capital endowment held constant, increasing the allowance "twisted" the lifetime earnings profile; increasing net earnings early in life but decreasing net earnings past birthday 45.

This model yields two results which provide a basis for estimating the ratio β_1/β_2 from data on individual students. The Cobb-Douglas coefficient on own-human capital is β_1 while the coefficient on purchased inputs is β_2. Data from a preliminary survey of the high school graduat-

[17]Confining observations to students at public four-year institutions minimizes the potential problem which might arise from students getting a larger allowance as the result of a choice to attend a more expensive institution. The allowance may vary among otherwise identical individuals in the same way as family income.

[18]Of course this finding may be an artifact of misspecification. With these data we may be observing some effects of nonhomogeneity in human capital and purchased inputs. Allowing for multiple dimensions in E and D may result in different production functions for E during school and on-the-job training. See Heckman for a discussion of problems of different production functions.

ing class of 1971 are used. Logit analysis of the labor force participation decision of 105 freshmen in state supported four-year colleges and universities indicates that the mean value of β_1/β_2 is between 6 and 7. Estimation of β_1/β_2 for a subsample of 40 working students yields a mean value of 2.38 with a standard error of 0.28, indicating that working and nonworking students may have different values of β_1/β_2. It is also shown that Haley's (1976) estimates of the returns to scale parameter in a model without purchased inputs is an unbiased estimate of $\beta_1 + \beta_2$ as surely as it is an unbiased estimate of his parameter. Combining Haley's estimate of β_1/β_2 with the estimates of β_1/β_2 can provide individual estimates of β_1 and β_2.

REFERENCES

Gary S. Becker, *Human Capital*, New York 1964.

———, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493–517.

———, "Human Capital and the Personal Distribution of Income: An Analytical Approach," in his *Human Capital*, 2d ed., New York 1975, 94–117.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

Merritt M. Chambers, *Higher Education and State Government, 1970–1975*, Danville 1974.

George F. Hadley and Murray C. Kemp, *Variational Methods in Economics*, New York 1971.

W. J. Haley, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–44.

———, "Estimation of the Earnings Profile from Optimal Human Capital Accumulation," *Econometricia*, Nov. 1976, *44*, 1223–38.

J. J. Heckman, "A Life-Cycle Model of Earnings, Learning, and Consumption," *J. Polit. Econ.*, Aug. 1976, *84*, 511–44.

Michael D. Intriligator, *Mathematical Optimization and Economic Theory*, Englewood Cliffs 1971.

T. Johnson, "Returns to Investment in Human Capital," *Amer. Econ. Rev.*, Sept. 1970, *60*, 546–60.

———, "Zealots and Malingerers: Results of Firm-Specific Human Capital Investments," *Southern Econ. J.*, Apr. 1975, *41*, 613–26.

E. Lazear, "Schooling as a Wage Depressant," work. paper no. 92, Center Econ. Anal. Hum. Behav. Soc. Instit., Nat. Bur. Econ. Res., Stanford, June 1975.

L. A. Lillard, "Human Capital Life Cycle of Earnings Models: A Specific Solution and Estimation," work. paper no. 4, Center Econ. Anal. Hum. Behav. Soc. Instit., Nat. Bur. Econ. Res., New York, July 1973.

———, "The Distribution of Earnings and Human Wealth in a Life Cycle Context," in F. Thomas Juster, ed., *The Distribution of Economic Well-Being*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 41, New York 1977.

D. McFadden, "Conditional Logic Analysis of Qualitative Choice Behavior," in Paul Zarembka, ed., *Frontiers in Econometrics*, New York 1973, 105–42.

D. O. Parsons, "The Cost of School Time, Foregone Earnings and Human Capital Formation," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, 251–66.

Karl Pearson, *Tables of the Incomplete Beta Function*, Cambridge 1934.

H. E. Ryder, F. P. Stafford, and P. E. Stephan, "Labor, Leisure and Training Over the Life Cycle," *Int. Econ. Rev.*, Oct. 1976, *17*, 651–74.

T. D. Wallace and L. A. Ihnen, "Full Time Schooling in Life Cycle Models of Human Capital Accumulation," *J. Polit. Econ.*, Feb. 1975, *83*, 137–55.

Research Triangle Institute, (1975a) "Preliminary National Longitudinal Study: Class of 1971," Contract OEC/O 6666 for the Center for Educational Statistics, Department of Health, Education, and Welfare, 1975.

———, (1975b) "National Longitudinal Study of the High School Class of 1972," Contract OEC/O 6666 for the Center for Educational Statistics, Department of Health, Education, and Welfare, 1975.

The Estimation of Labor Supply Models Using Experimental Data

By MICHAEL C. KEELEY, PHILIP K. ROBINS, ROBERT G. SPIEGELMAN, AND RICHARD W. WEST*

For many years there has been interest in replacing the existing complex transfer system in the United States with a nationwide negative income tax (*NIT*) program.[1] The feasibility and desirability of an *NIT*, however, depend on its effects on aggregate labor supply (and its cost). Interest in predicting these aggregate effects has motivated considerable empirical research on labor supply. The first studies used existing data, usually cross-sectional, to estimate the parameters of labor supply functions.[2] Unfortunately, the range of estimates in these studies is disturbingly large and of limited usefulness to policymakers.[3] Consequently, a new approach to labor supply research has been followed - social experimentation.[4]

Several experiments have been funded by the federal government to test the effects of alternative *NIT* programs on labor supply. The first experiment, the New Jersey Experiment, was conducted in New Jersey and Pennsylvania from 1968 to 1972.[5] Other experiments have taken place in Gary, Indiana from 1970 to 1974, and in rural areas of Iowa and North Carolina from 1969 to 1973. The largest and most comprehensive of these experiments began in 1971 in Seattle, Washington and Denver, Colorado and is still taking place.

In principle, a controlled experiment affords the opportunity to overcome most of the problems inherent in nonexperimental research, because in an experiment, the budget constraints of individuals are exogenously shifted in a measurable way. In practice, however, the experiments have been beset with their own unique set of econometric problems. These problems include the nonrandom assignment of experimental treatment, small samples, truncation of response, limited duration, participation in other welfare programs both before and during the experiment by sample members, and the selection of nonrepresentative samples.[6]

In this paper, a methodology is presented that attempts to deal with these problems. Experimental data from the Seattle and

*Economists, SRI International. The research reported in this paper was performed under contracts with the states of Washington and Colorado, prime contractors for the Department of Health, Education, and Welfare, under contract numbers SRS-70-53 and SRS-71-18, respectively The opinions expressed in the paper are our own and should not be construed as representing the opinions or policies of the states of Washington or Colorado, or any agency of the *U S* government. An earlier version of this paper was presented at the Summer 1976 meetings of the Econometric Society and in seminars at the National Bureau of Economic Research and Mathematica Policy Research. Jodie Allen, Yoram Barzel, David Betson, Michael Boskin, Glen Cain, Joseph Corbett, Irwin Garfinkel, David Greenberg, Terry Johnson, Richard Kaluzny, Richard Kasten, Robert Lerman, Stanley Masters, Myles Maxfield, Robert Moffit, Larry Orr, Harold Watts, and Robert Willis provided valuable comments on various drafts of this paper. We are, of course, solely responsible for the views presented and for any remaining errors. Helen Cohn, Diane Hollenbeck, Paul McElherne, Gary Stieger, and Steven Spickard provided expert programming assistance.

[1]Milton Friedman is usually credited with developing the concept of a negative income tax. Robert Lampman and James Tobin (1965) among others also made early contributions to the concept.

[2]An excellent collection of such studies is presented in Glen Cain and Harold Watts.

[3]See Keeley for a survey of these studies and a discussion of some of the econometric difficulties that lead to such a wide range of estimates.

[4]Heather Ross (1966) is credited with first conceiving the idea of an *NIT* experiment Guy Orcutt and Alice Orcutt (1968) first published a paper outlining an experimental design.

[5]The New Jersey Experiment is described in David Kershaw and Jerilyn Fair. Watts and Albert Rees (1977a, b) and Joseph Pechman and P. Michael Timpane present the results from this experiment.

[6]See Henry Aaron, Keeley, and Keeley and Robins for a critical discussion of many of these problems.

Denver Income Maintenance Experiments (*SIME/DIME*) are used to estimate the parameters of a labor supply function.[7] These parameters are then used to predict the nationwide labor supply effects of alternative *NIT* programs.

The empirical response function estimated measures the change in labor supply over a two-year period. The Tobit method is used to estimate equations for single female heads of families, husbands, and wives. Nationwide aggregate labor supply responses to six alternative *NIT* programs are predicted by applying the response function to data from the March 1975 *Current Population Survey*.

The plan of the paper is as follows: Section I describes an experimental *NIT* program; Section II presents a theoretical model of the labor supply response to an *NIT*; Section III specifies the empirical model; Section IV presents the empirical results; Section V discusses policy implications; and Section VI presents the summary and conclusions.

I. Description of an Experimental *NIT* Program

An *NIT* program is characterized by a support (or guarantee) level S, and a tax rate t_e. The support level is the grant provided when other income is zero, and the tax rate is the rate at which the grant declines as other income increases. In a controlled *NIT* experiment, an effort is made to ensure that the influence of other tax and transfer programs is eliminated. Public transfers, therefore, are fully taxed, and positive taxes are reimbursed. Consequently, the payment a person receives depends on gross income and both experimental and nonexperimental tax rates.

For this discussion, it is assumed that nonexperimental net nonwage income (including public transfers) is zero and that both the nonexperimental and experimental tax rates are constant. These assumptions are relaxed in the empirical analysis. The payment P associated with a particular *NIT* program is determined as follows:

$$(1)\quad P = \begin{cases} S - t_e Y + t_n Y & \text{if } S + t_n Y \geq t_e Y \\ 0 & \text{if } S + t_n Y < t_e Y \end{cases}$$

where t_n is the nonexperimental tax rate and Y is gross income. The payment, if positive, is equal to the grant $S - t_e Y$, plus the positive tax reimbursement $t_n Y$.

Figure 1 shows a graph of the nonexperimental budget line (line ABT) and the experimental budget line (line ABE) of an individual enrolled in an *NIT* program. Point B, where the two budget lines intersect, is the point at which the payment becomes zero and is known as the tax break-even level. The tax break-even level of income, given by $S/(t_e - t_n)$, may be contrasted with the grant break-even level of income, given by S/t_e. The grant break-even level is the level of income at which the *NIT* grant $(S - t_e Y)$ becomes zero.

The program support level S is designated by the line ET. Under the assumption that the nonexperimental tax rate, t_n, is less than the tax rate of the *NIT* program under consideration, t_e, the absolute value of the slope of the new budget line (ABE) to the right of B is reduced.

The Seattle and Denver Income Maintenance Experiments are testing eleven different *NIT* programs. The programs are described in Table 1. A feature of *SIME/DIME* that distinguishes it from the other *NIT* experiments is the testing of programs in which the marginal tax rate declines as income increases. Families in *SIME/DIME* are enrolled for either three or five years.[8] Different durations are being tested, because of difficulties in inferring permanent effects from experiments of finite length. According to Charles Metcalf, substitution and income effects should vary according to the length of the experiment. In our empirical analysis, we formally test for such differences. The results of these tests are presented in Section IV.

[7] For a description of *SIME/DIME*, see Mordecai Kurz and Spiegelman (1971, 1972).

[8] A small number of families are enrolled for twenty years but are not considered in this study.

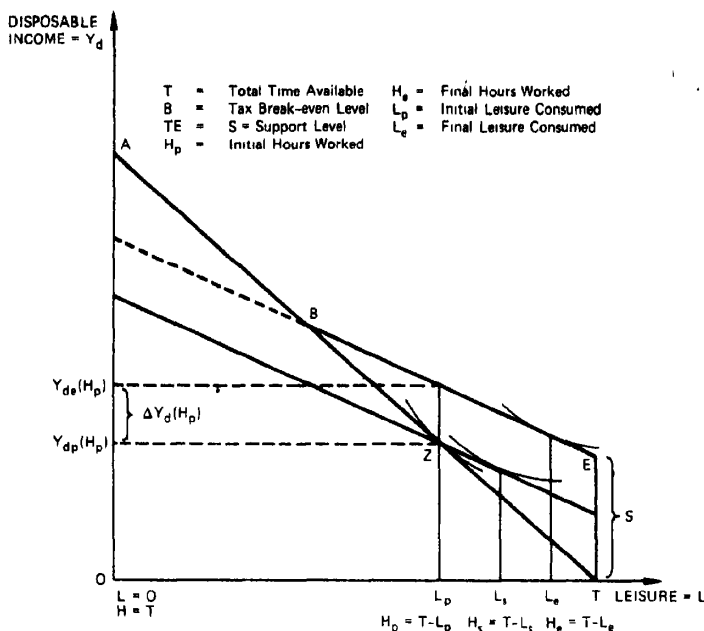



FIGURE 1. AN EXPERIMENTAL *NIT* PROGRAM

TABLE 1—PROGRAMS BEING TESTED IN THE SEATTLE AND DENVER INCOME MAINTENANCE EXPERIMENTS
(1971 Dollars)

Support Level	Initial Tax Rate	Rate of Decline of Average Tax Rate per $1,000 of Income	Grant Break-Even Level	Tax Break-Even Level
$3,800	.5	0	$ 7,600	$10,250
3,800	.7	0	5,429	6,350
3,800	.7	.025	7,367	10,850
3,800	.8	.025	5,802	7,800
4,800	.5	0	9,600	13,150
4,800	.7	0	6,867	8,520
4,800	.7	.025	12,000	19,700
4,800	.8	.025	8,000	11,510
5,600	.5	0	11,200	15,700
5,600	.7	0	8,000	9,780
5,600	.8	.025	10,360	16,230

Note: The figures for the support level, the grant break-even level, and the tax break-even level are in 1971 dollars and are for a family of four with only one earner and no income outside of earnings. Adjustments are made to these figures for family size and for changes in the cost of living over time. Positive tax reimbursements include the federal income tax and social security taxes. The federal income tax assumes the family takes the standard deduction. State income taxes, which are relevant only for the Denver Experiment (there is no state income tax in Washington), are ignored in calculating the tax break-even level. Thus, the tax break-even level is slightly higher for the Denver Experiment.

II. Theoretical Analysis of the Labor Supply Response to an *NIT* Program

A. *The Model*

It is assumed that each individual maximizes a well-behaved utility function, $U(L, Y_d)$, where L is leisure and Y_d is consumption of market goods (or disposable income) subject to the budget constraint

$$F \equiv wT + Y_n = wL + Y_d \tag{2}$$

where F is full income, w is the net wage rate, T is total time available, and Y_n is net nonwage income. Utility maximization implies that the individual has a labor supply function $H = H(w, Y_n)$, where $H = T - L$ is hours of work. Totally differentiating the labor supply function and substituting in the Slutsky equation[9] gives

$$\begin{aligned}(3)\quad dH &= \left.\frac{\partial H}{\partial w}\right|_U \cdot dw + \frac{\partial H}{\partial Y_n}(Hdw + dY_n)\\ &= \alpha dw + \beta(Hdw + dY_n)\end{aligned}$$

where U is utility, α is the substitution effect, and β is the income effect.[10] The term $Hdw + dY_n$ is the total differential of disposable income, holding constant the initial supply of labor H.

The model given by equation (3) is specified in terms of unobservable differential changes. The differential change model implies that each individual's point of compensation should be his or her initial equilibrium labor supply. If differences in initial labor supply across individuals are the result of differences in equilibrium or permanent labor supply, each person should be compensated at his or her initial position. In our application of the model, we follow this compensation procedure.[11] To measure substitution and income effects empirically, finite differences are used to approximate the unobservable differential changes. In discrete form, the model described in equation (3) becomes:

$$\begin{aligned}(4)\quad \Delta H \approx \alpha\Delta w + \beta(H_p\Delta w + \Delta Y_n) =\\ \alpha\Delta w + \beta\Delta Y_d(H_p)\end{aligned}$$

where $\Delta Y_d(H_p)$ is the change in disposable income of an individual, holding constant his or her initial labor supply H_p.

B. *Analyzing the Response to an NIT Program*

Equation (4) states that the effects of shifts in the budget constraint on labor supply can be decomposed into a substitution effect, which depends on the change in the net wage rate Δw, and an income effect, which depends on the change in disposable income evaluated at initial hours of work, $\Delta Y_d(H_p)$. For a person placed on an experimental *NIT* program, Δw is equal to the gross wage rate times the difference between the preexperimental and experimental tax rates,[12] and $\Delta Y_d(H_p)$ is equal to the payment the person would receive if initial labor supply were maintained. Referring again to Figure 1, consider a person below the break-even level who is in equilibrium at point Z before the imposition of an *NIT* program. The change in the quantity of leisure demanded is comprised of a substitution effect ($L_s - L_p$) holding disposable income constant at initial labor supply, and an income effect ($L_e - L_s$) holding relative prices (i.e., the

[9]The Slutsky equation decomposes the total effect of a wage change on labor supply into a substitution effect and an income effect:

$$\frac{\partial H}{\partial w} = \left.\frac{\partial H}{\partial w}\right|_U + H\frac{\partial H}{\partial Y_n}$$

[10]For a family with more than one potential earner, the equation can be generalized to include cross-substitution effects. In our empirical formulation of this model, it is assumed that cross-substitution effects are zero, partly because the net wage changes of both spouses are highly correlated and their effects are difficult to distinguish empirically. An attempt to apply this model to nonexperimental cross-sectional data is presented in Orley Ashenfelter and James Heckman (1973, 1974). See, however, the critique of Jonathan Dickinson (1977).

[11]If differences in initial labor supply are purely transitory, then such a procedure is not appropriate, because initial labor supply is not at an equilibrium position. Compensation at the initial position, however, ensures that the point of compensation is not endogenous.

[12]This assumes that the gross wage rate is unaffected by the program.

wage rate relative to the price of goods) constant.[13] Disposable income is held constant by rotating the budget line through the initial equilibrium point Z, where, at the new net wage rate and new monetary full income, the consumer could still purchase the initial consumption bundle.

The analysis thus far focuses on the response of a given individual to a particular program. Because of differences in tastes or other unmeasured variables, however, there is considerable heterogeneity of the initial equilibrium positions of individuals.[14] In fact, the best empirical labor supply equations explain only about 20–30 percent of the variance in labor supply.[15] This suggests that on a given budget line there is a distribution of initial equilibrium positions. For simplicity, it is assumed that this distribution is the result of differences in tastes.

Because separate responses resulting from compensated wage changes and income changes are not observed for each person (only total response that results from both changes is observed),[16] some a priori restriction is needed to identify the model so that income and substitution effects can be measured empirically. The restriction we impose is to assume that different individuals have equal substitution effects and equal income effects at their initial equilibrium positions.[17] If, in fact, income and substitution effects differ among individuals, the empirical method used measures average income and substitution effects in the sample.

The assumptions underlying this model are different from those implicit in most cross-sectional studies.[18] Instead of assuming that each person has the same preference structure, it is assumed that differences in taste are reflected in differences in initial equilibrium labor supply, after controlling for differences in budget constraints. Therefore, there is no single utility function that is consistent with the model, although each person is assumed to maximize a well-behaved utility function.

C. *Implications of the Model*

The assumption that different individuals have equal substitution and income effects implies that response to a given *NIT* program depends on the initial equilibrium position. For example, a person with low income experiences a considerable change in disposable income and net wage rate, and a large response is expected. On the other hand, a person initially at the break-even level experiences a change only in the net wage rate. Response for this person consists only of a (Slutsky) substitution effect and is therefore smaller. Next, consider a person above the break-even level. This person experiences changes in disposable income and the net wage rate only if the elasticity of substitution in consumption is sufficiently large that the indifference curve through the initial point intersects the *NIT* segment of the new budget line. Thus, for a person initially above the break-even level, we would expect a very small probability of response. Finally, consider an individual who is not working: this person experiences a considerable change in disposable income and net wage rate, but has zero response.[19]

[13]This is the Slutsky, as opposed to the Hicks, decomposition.

[14]See Robert Hall (1975), and Heckman and Robert Willis for a discussion of heterogeneity.

[15]See Cain and Watts for a sampling of typical cross-section labor supply equations. These studies, however, analyze a measure of labor supply that does not correspond strictly to our concept of equilibrium labor supply. Instead, the studies use current labor supply, which is the sum of permanent or equilibrium labor supply, a transitory component, and a life cycle component.

[16]For persons at the tax break-even level initially, total response is a result of the (Slutsky) substitution effect.

[17]Although it may appear that we are assuming constant substitution and income effects for each person, this is not the case; indeed, it is impossible to have a labor supply function with constant income and substitution effects (see, for example, Dickinson, 1975, p. 31).

[18]It might be noted that the model described in this paper cannot be estimated using cross-sectional data. In a cross section, only one equilibrium position is observed, and the model is not identified.

[19]Response is subject to truncation because, at most, a person can reduce hours to zero. The estimation technique we employ accounts for this problem.

Thus, response depends on the initial equilibrium position.[20]

Differences in response to an *NIT* program arise, not because persons with different tastes for work have inherently different responses to changes in disposable income or net wage rates, but because individuals with different propensities to work (different initial equilibria on a given budget constraint) are offered different inducements to change their behavior. Those with the smallest propensities to work experience the largest changes in income.

A final implication of this model is that theory-free response models that compare the average response of persons on different programs are not meaningful. The reason for this is that persons with higher incomes (and therefore higher labor supplies) are assigned to the more generous programs in order to reduce the average cost of an observation.[21] Thus, because both response and assignment to program depend on the initial position, biased measures of program differences are obtained.[22] The response model used, however, controls for the nonrandom assignment by allowing response to be a function of preexperimental labor supply and by directly measuring the change in budget constraints caused by the *NIT*.

III. Empirical Specification

To estimate equation (4), data on heads of families in *SIME/DIME* are used. The change in labor supply ΔH, is equal to hours of work in the second year of the experiment H_e, minus hours of work in the year prior to the experiment H_p. The response variables, Δw and $\Delta Y_d(H_p)$, depend on the particular budget constraint, and on the preprogram equilibrium position. For reasons described below, several modifications are made to this equation regarding functional form, additional variables, missing data on wage rates for nonworkers, and nonlinearity of the budget constraints.

A. *The Role of Control Families*

Approximately 45 percent of the families in *SIME/DIME* serve as controls and are not eligible for payments. For these families, it is assumed that Δw and $\Delta Y_d(H_p)$ are zero.[23] Control families are included in the sample, however, to increase the efficiency of the estimated treatment effects.[24] Efficiency is increased because factors other than the experiment (such as changing economic conditions) cause labor supply to change over time. The inclusion of control families in the sample enables us to make a more precise distinction between experimental and nonexperimental effects. Variables used in this study to measure nonexperimental effects are called control variables. The control variables include all variables that affect assignment to experimental treatments.[25]

B. *Calculating the Change in Disposable Income*

The change in disposable income evaluated at the initial equilibrium hours of work

[20] Note that in a typical cross-sectional model, where it is assumed that gross wage effects and income effects are constant, response would not depend on the initial position (below break even and ignoring truncation), because the changes in nonwage income and the net wage rate do not depend on the initial position.

[21] Simple random assignment is not used in any of the *NIT* experiments. See John Conlisk and Kurz, and Keeley and Robins for a description of the *SIME/DIME* assignment model.

[22] If program dummy variables were interacted with all assignment variables, unbiased estimates of response could be obtained. Such a model would have far too many parameters, however, to be estimated with precision using our sample. See Spiegelman and West.

[23] Changes in the control budget constraints that have zero mean and are uncorrelated with the variables in the equation would not affect the consistency of the estimates.

[24] A comparison of least squares estimates for husbands, including and excluding control families, indicates that the coefficients of $\Delta Y_d(H_p)$ and Δw differ by less than 10 percent, while the standard errors are 16 percent larger when controls are excluded.

[25] The control variables include eight dummy variables for normal income categories, dummy variables for race (black/white) and site (Seattle/Denver), age, number of family members, number of children under 5 years of age, and Aid to Families with Dependent Children (*AFDC*) benefits in the year prior to enrollment.

is given by

$$(5) \quad \Delta Y_d(H_p) = Y_{de}(H_p) - Y_{dp}(H_p)$$

where $Y_{de}(H_p)$ is disposable income evaluated at H_p under the *NIT*, and $Y_{dp}(H_p)$ is disposable income evaluated at H_p before the *NIT*. For this study, $\Delta Y_d(H_p)$ is calculated on the basis of earnings and nonwage income in the year before enrollment in the experiment.[26] Thus, $\Delta Y_d(H_p)$ depends on both transitory and permanent components of labor supply. In theory, the change in disposable income should be measured at normal or permanent labor supply. Because there is likely to be a transitory component in our measure of labor supply, our estimate of the income effect will be biased because of the presence of errors in variables.[27] In a lengthier version of this paper available from the authors upon request, the bias is discussed and it is shown that the bias is not likely to be large if preexperimental labor supply is included on the right-hand side of the equation. For this reason, H_p is included among the explanatory variables, and H_e is used as the dependent variable.

C. Calculating the Change in the Net Wage Rate

The change in the net wage rate is given by

$$(6) \quad \Delta w = -W(t_e - t_p)$$

where W is the gross wage rate, t_e is the experimental tax rate, and t_p is the preexperimental tax rate. In calculating this variable from *SIME/DIME* data, two problems arise. First, the preexperimental tax function and many of the experimental tax functions are non-linear. Second, wage rates are not observed for nonworkers.

As mentioned earlier, a feature of *SIME/DIME* is the testing of declining tax rate programs. The effect of the declining tax rate programs is to make the experimental tax rate t_e an endogenous variable that depends on labor supply. To purge the tax rate of this endogeneity, we linearize the budget constraint around the preexperimental point and treat the individual as if he or she were on the tangent linear budget constraint. This procedure is deficient in that all final equilibrium points are not on the linearized budget constraint, although the rate of decline of the tax rate is small. Furthermore, because the experimental budget set is nonconvex for families on the declining tax rate programs, and because small changes in nonconvex budget sets may lead to large changes in behavior, the linearization may not be a reasonable approximation to the true budget set. To account for the linearization procedure during estimation, we include a dummy variable for persons on the declining tax rate programs.

The preexperimental budget constraint is also non-linear, because of the progressivity of the positive income tax system and the interrelations among tax rates in income-conditioned public transfer programs. Endogeneity is not a problem, however, because preexperimental labor supply is predetermined. The preexperimental tax rates are derived on the basis of preexperimental

[26] $\Delta Y_d(H_p) = S - SR100 - .5(SR50 - SA50) - [t - r(Y - E)](Y - E) + Q$, where S is the support level, t is the initial tax rate, and r is the rate of decline of the average tax rate The term $SR100$ represents items taxed at 100 percent. bonus value of food stamps, welfare benefits other than *AFDC*, unemployment and workmen's compensation, veteran's survivors and disability benefits, training stipends net of tuition, fees and books, and social security benefits. The term $SR50$ represents items taxes at 50 percent: alimony and child support received and other support received The term $SA50$ represents items reimbursed at 50 percent. alimony and child support paid and other support paid. The term Y represents items taxed as income· earnings, insurance benefits, pensions and annuities, payments from private disability plans, and a fraction of net worth. The term E represents items subtracted from income· child care expenses, care for the aged, and medical expenses. The term Q represents items reimbursed at 100 percent: federal and state income taxes and social security taxes. If $\Delta Y_d(H_p) > 0$, a family is defined as being below the tax break-even level; it is set equal to zero for families above the tax break-even level. Families receiving *AFDC* benefits prior to the experiment are required to give up their *AFDC* status in order to receive *NIT* payments. We subtract preexperimental *AFDC* benefits from $\Delta Y_d(H_p)$ for families below the tax break-even level.

[27] The substitution effect would also be biased to the extent that the preexperimental tax rate depends on preexperimental labor supply.

income and participation in certain income-conditioned tax and transfer programs. The income-conditioned programs we consider include federal and state income taxes, social security taxes, *AFDC*, Aid to Families with Dependent Children-Unemployed Parent (*AFDC-UP*), and Food Stamps. The tax rates are derived in accordance with the tax laws and the administrative regulations of the public transfer programs.[28]

Because wage rates are not observed for nonworkers, a wage equation is estimated for workers based on personal characteristics, and the equation is used to predict wage rates for the entire sample.[29] A variety of different wage equations can be specified; the wage equation we estimate is a simple linear formulation based on the human capital model of Jacob Mincer.[30] The change in the net wage Δw is calculated as the product of the predicted wage rate and the difference between the linearized preexperimental and experimental tax rates. Like $\Delta Y_d(H_p)$, Δw is set equal to zero for persons above the tax break-even level.

D. *Additional Experimental Variables*

Certain families on *SIME/DIME* are enrolled in manpower programs that provide counseling and subsidize training and educational activities.[31] To capture the effects of the three manpower programs of the experiment, dummy variables for each program are included in the empirical specification.

Many of the enrolled families are initially above the tax break-even level.[32] Even though the calculated values of Δw and $\Delta Y_d(H_p)$ are zero for families with preexperimental equilibria above the tax break-even level, some of these families will respond to the experiment.[33] Response above the break-even level is measured by defining three explanatory variables that capture the location of the family relative to the break-even level: a dummy variable signifying whether or not the family is above the break-even level, the break-even level of the

[28] See Kurz et al. for a discussion of how the positive tax rates are derived, and Maxfield for a discussion of how the transfer program tax rates are derived. There is some evidence that legal tax rates are an overestimate of the effective tax rates of public transfer programs. Legal tax rates are used in this paper because they are used in the computer program that extrapolates the experimental results to the national population.

[29] This approach follows Hall (1973) and Edward Kalachek and Frederick Raines. Reuben Gronau and Heckman (1974) demonstrate that the wage equation approach yields biased estimates for nonworkers, and they develop alternative estimation procedures. However, in this paper, the substitution effect is estimated as the coefficient of the change in the net wage rather than the coefficient of the gross wage rate It is unlikely that small biases in estimating gross wages significantly affect the change variable, which depends primarily on the difference between the experimental and preexperimental tax rates. In a recent paper, Heckman (1976) finds that in a national sample of white married women (the National Longitudinal Survey) the selectivity bias in wage rates is quantitatively small.

[30] The estimated wage equations are

$$W = -.071B + .033E + .061X - .00106X^2 \quad + 2.340 \text{ for husbands}$$
$$\quad\;\; (.051) \quad (.012) \quad (.008) \quad (.00018) \qquad\qquad (.172)$$

$$W = .800B + .110E + .036X - .00073X^2 \quad + .590 \text{ for wives}$$
$$\quad\;\; (.049) \quad (.014) \quad (.008) \quad (.00020) \qquad\qquad (.192)$$

$$W = .010B + .102E + .045X - .00098X^2 \quad + .816 \text{ for female heads of households}$$
$$\quad\;\; (.045) \quad (.013) \quad (.009) \quad (.00021) \qquad\qquad (.183)$$

where B is a dummy variable for race (black = 1), E is years of schooling, and X is experience (defined as age minus years of schooling minus 5). Standard errors are in parentheses. The R^2s are .112, .048, and .116, respectively Because the variance of Δw would be dominated by the change in the tax rate no matter how complicated the wage equation, the results using a more complicated wage equation are likely to be similar to the results reported in this paper.

[31] See Kurz and Spiegelman (1971, 1972) for a description of the manpower component of *SIME/DIME*

[32] Based on preexperimental income, 10 percent of the single-parent headed families and 20 percent of the double-parent headed families in *SIME/DIME* are above the tax break-even level.

[33] Under the assumption that substitution effects are constant, it can be shown that families above the tax break-even level will respond only if income in excess of the break-even level is less than half of the absolute value of the change in income they would experience if they did respond (see Robins and West).

family earnings, and the amount of family earnings above the break-even level.[34]

E. *Estimation Procedure*

Because H_e cannot take on negative values and because there are numerous observations where $H_e = 0$, estimation of the model by ordinary least squares would yield inconsistent coefficient estimates. Furthermore, the estimates would be inefficient because the error term is heteroscedastic. To account for these statistical problems, we use a tobit model, which is designed to handle cases where the dependent variable is truncated normal.[35] The Tobit model may be written as

$$(7)\quad H_e = \max[b_0 + b_1 H_p + b_2 C + b_3 M + b_4 \Delta Y_d(H_p) + b_5 \Delta w + b_6 FABOVE + b_7 BREAK + b_8 EARNABV + b_9 DECLINE + e, 0]$$

where
H_e = experimental hours of work
H_p = preexperimental hours of work
C = vector of control variables
M = vector of manpower treatment variables
$\Delta Y_d(H_p)$ = change in disposable income evaluated at preexperimental labor supply (thousands of dollars per year)
Δw = change in the net wage rate (dollars per hour)
$FABOVE$ = dummy variable for persons above the tax break-even level
$BREAK$ = break-even level of earnings (thousands of dollars per year)
$EARNABV$ = family earnings above the break-even level (thousands of dollars per year)
$DECLINE$ = dummy variable for persons on the declining tax rate programs.
e = random error term, assumed to be distributed normally with variance σ^2.

The b_i and σ^2 are estimated by maximum likelihood using an iterative-maximization technique.

The parameters in a Tobit model cannot be interpreted in the same way as the parameters in a linear model. In a linear model, the treatment parameters are interpreted as the average response of the population to the imposition of a negative income tax. In a Tobit model, the treatment parameters give the average response only of persons who have nonzero labor supplies (interior solutions) before and after the imposition of a negative income tax. The response of all other persons is somewhat smaller in magnitude than that of persons with interior solutions because of the lower bound on the dependent variable. The coefficients of $\Delta Y_d(H_p)$ and Δw, however, can be interpreted as income and substitution effects for persons with interior solutions before and after the *NIT* program is implemented.

This empirical specification eliminates many of the problems associated with measuring the response to an *NIT* program. Nonrandom assignment by family income is taken into account because the response is allowed to vary with preexperimental income, which is the major assignment variable. Heterogeneity is partially controlled because individuals with identical budget

[34]Our specification of these above break-even variables is likely to suffer from errors of measurement of the same type as those present in $\Delta Y_d(H_p)$. To some extent, however, the procedure used to account for errors of measurement in $\Delta Y_d(H_p)$ should also account for this type of measurement error. The primary cause of bias in the above break-even variables is probably misclassification of persons who are near the break-even level preexperimentally; the specification would thus lead to overestimation of effects above the break-even level and underestimation of effects below the break-even level. In another paper, Robins and West present a model that unifies the response above and below the break-even level and find that the results are similar to those presented in this paper.

[35]See Takeshi Amemiya or Tobin (1958) for a discussion of the Tobit model.

TABLE 2—ESTIMATED EXPERIMENTAL EFFECTS ON LABOR SUPPLY
(Tobit Estimates)

	Coefficient		
Independent Variable	Husbands	Wives	Female Heads
Below break-even			
$\Delta Y_d(H_p)$	−34.4	−142.9[c]	−101.1[b]
	(27.3)	(44.4)	(39.4)
Δw	83.2[b]	168.0[a]	125.8[a]
	(37.1)	(91.2)	(65.9)
Above break-even			
FABOVE	−12.7	−430.8[a]	−344.8
	(174.6)	(255.6)	(291.3)
BREAK	−5.5	8.3	73.2
	(21.1)	(29.5)	(64.7)
EARNABV	11.5	47.5	35.1
	(27.3)	(42.0)	(55.6)
DECLINE	−86.3[b]	119.5	21.8
	(48.4)	(78.1)	(73.2)
χ^2	21.55[c]	26.84[c]	20.24[c]
S	720	1,086	990
	(14)	(28)	(25)
$\bar{H}_e$	1,736	659	975
	(825)	(825)	(935)
N	1,592	1,698	1,358

Notes: Standard errors in parentheses, χ^2 is the *chi*-square test for treatment effects (6 degrees of freedom); S is the standard error of estimate; $\bar{H}_e$ is the mean of the dependent variable, hours of work per year in the second year of the experiment, N is the sample size.
[a]Indicates significance at 10 percent level.
[b]Indicates significance at 5 percent level
[c]Indicates significance at 1 percent level.

constraints are allowed to respond differently. Preexperimental participation in other welfare programs is taken into account by including welfare income and tax rates in the definitions of the changes in disposable income and net wage rates. Finally, the estimation of substitution and income effects enables the prediction of labor supply response to *NIT* programs other than the ones being tested in *SIME/DIME*.

IV. Results

The sample consists of a subset of originally enrolled black and white family heads who remained in *SIME/DIME* for at least two years and for whom data were available at the time this study was undertaken.[36] The empirical model is estimated separately for female heads of households and for husbands and wives in two-parent headed households. The subgroups for analysis are defined as of the date of enrollment regardless of changes in marital status. This approach is used so that the estimates are not conditional on unchanged marital status.

[36]About 800 Mexican-Americans are enrolled in the Denver Experiment but are excluded from the analysis in this paper because data were not available for them when this study was undertaken. It has turned out to be a rather difficult task to build computer software that converts data from interview form into analytical files with reasonable flexibility and generality at low cost. SRI International is now in the process of building such computer software and processing interview data into a data management system. This system will enable users to construct their own analytical files from the basic data. When this data base system is complete, the Department of Health, Education, and Welfare, which is funding *SIME/DIME*, will release a public use file. In the interim, a copy of the data tape used for the analysis presented in this paper is available on request (at a nominal cost to cover copying and documentation).

TABLE 3—TESTS OF SITE, RACE, AND EXPERIMENTAL DURATION DIFFERENCES IN RESPONSE

	Husbands	Wives	Female Heads
Site test	.53	.41	1.12
Race test	1.70	.52	1.34
Experimental duration test	.93	1.53	.41

Notes. Tests are based on ordinary least squares estimates. The race and site tests are performed by interacting each experimental variable with race and site dummies. The experimental duration test is performed by interacting Δw and $\Delta Y_d(H_p)$ with dummy variables for the three- and five-year programs. The coefficients of control variables are constrained to be the same in the tests. Numbers given are F-ratios with 6 and N degrees of freedom for the race and site tests, and 2 and N degrees of freedom for the experimental duration tests, where N is the sample size.

Table 2 displays the Tobit estimates for the experimental variables.[37] In Table 3, the results of tests of differences in response by race, experimental site, and experimental duration are presented. The various tests are performed using ordinary least squares to reduce computational expense.

For each group, there are statistically significant experimental effects on labor supply. The income effects are negative and statistically significant for wives and female heads, and the substitution effects are positive and statistically significant for all three groups. The F-statistics for site and race differences are not significant, implying that the hypotheses of equal experimental effects in Seattle and Denver and for blacks and whites cannot be rejected. The tests of different substitution and income effects for persons on the three- and five-year experimental programs are not statistically significant; however, for all three groups, the substitution effect is larger and the income effect is smaller for persons on the three-year programs, a result consistent with the predictions of the model developed by Metcalf.[38] For families above the tax break-even level, only wives appear to be responding to the experiment. All three groups exhibit a response above the break-even level that declines in absolute value with distance from the break-even level.[39]

Table 4 presents estimated substitution and income effects evaluated at the sample means for persons below the break-even level who are working in both the pre-experimental and experimental periods (for example, persons for whom $H_e > 0$ and $H_p > 0$).

The estimated effects at the sample means

[37]The results for the control and manpower variables are in an appendix available upon request from the authors.

[38]*SIME/DIME* is uniquely structured to test the effects of experimental duration on behavioral response. Studies devoted to this issue are currently being undertaken.

[39]Because the functional form used for the experimental response could be considered fairly restrictive, we have estimated several less restricted versions of the model. These versions include the following sets of additional variables; 1) a dummy variable for having a financial treatment, 2) the change in the support level, 3) interactions of $\Delta Y_d(H_p)$ with preexperimental nonwelfare income, and 4) interactions of Δw and $\Delta Y_d(H_p)$ with *DECLINE*. Out of twelve tests of the null hypotheses that these additional variables have zero coefficients, only one, the test of 1) for wives, is significant at the 5 percent level; all the others are not significant at the 10 percent level.

TABLE 4—SUBSTITUTION AND INCOME EFFECTS AT THE MEAN (Estimated Asymptotic Standard Errors in Parentheses)

	Husbands	Wives	Female Heads
Substitution effect at the mean ($\hat{b}_5\Delta\overline{w}$)	-55.7	-63.8	-59.1
	(24.9)	(34.7)	(31.0)
Income effect at the mean $[\hat{b}_4\Delta Y_d(H_p)]$	-47.1	-198.6	-117.3
	(37.4)	(61.7)	(45.7)
Total effect at the mean	-102.8	-262.4	-176.4
	(33.0)	(55.1)	(43.6)
Mean hours of work in preexperimental period ($\overline{H}_p$)	1,922	1,194	1,577

indicate a substantial disincentive effect, particularly for women. In percentage terms, the effects are −5.3 percent for husbands, −22.0 percent for wives, and −11.2 percent for female heads of families. It is important to note that these effects are based on mean changes in disposable income and net wage rates that result from the set of programs being tested in *SIME/DIME*, rather than from any single *NIT* program.

V. Implications of the Results for a Nationwide *NIT* Program

One of the primary reasons for undertaking the *NIT* experiments is to provide policymakers with estimates of the labor supply effects of a nationwide *NIT* program. The model developed in this paper can be used to predict the labor supply effects of a variety of nationwide *NIT* programs, including programs that are different from those being tested in the experiments.

We use the March 1975 *Current Population Survey* (*CPS*) to generate nationwide predictions.[40] The *CPS* is a weighted random sample of the *U.S.* population and contains information on about 50,000 households.[41] The predictions are derived by applying the estimated response function to each individual and then summing the estimated responses over all individuals. Only the responses of heads of families between the ages of 18 and 58 are considered; nonheads of households, households with only one member, and the elderly are omitted from the analysis.

Predictions are generated for six different *NIT* programs. The six programs have constant tax rates of 50 and 70 percent and support (guarantee) levels of 50, 75, and 100 percent of the poverty level ($5,000 for a family of four in 1974). Because the poverty level increases with family size, the support level also increases with family size. The nominal support level is constant across regions, and the *NIT* program is assumed to replace the existing *AFDC*, *AFDC-UP*, and Food Stamp programs. All other nonlabor income is taxed by the program at a rate of 100 percent.

The predicted labor supply responses are presented in Table 5 and are reported in two ways: first, the average responses for all participating families; and second, the average responses for the *U.S.* population. The average responses for the *U.S.* population include the responses of certain nonparticipants, as well as the responses of participants. The nonparticipants who respond are families that previously received welfare benefits and are above the break-even level of the *NIT* program. These families increase their labor supply when the welfare programs are replaced by the *NIT* program.

In interpreting the results, it is important to keep in mind that the responses vary not only because of changing guarantee levels and tax rates, but also because of a changing pool of participants. For example, as the tax rate increases (for a given guarantee), the pool of participants decreases. On the other hand, as the guarantee increases (for a given tax rate), the pool of participants increases. The manner in which the pools change depends on the distribution of income within the relevant population subgroup.

For participating husband-wife families, the magnitudes of the average responses are positively associated with both the guarantee and the tax rate. For participating female-headed families, the responses are positively associated with the guarantee, but do not vary with the tax rate. For both groups, the results indicate fairly sizable reductions in labor supply, ranging from between 10 and 21 percent for husband-wife families and between 0 and 15 percent for female-headed families.

The average responses of the *U.S.* population are quite small relative to the average responses of participating families because most families in the United States do not choose to participate in the program. While the magnitudes of the average responses again increase with the guarantee (as they

[40] For a detailed description of the methodology used to generate the predictions, see Keeley et al. and Maxfield.

[41] The income data from the March 1975 *CPS* are annual data for the year 1974.

TABLE 5—AVERAGE LABOR SUPPLY RESPONSES TO A NATIONWIDE *NIT* PROGRAM FOR ALL PARTICIPATING FAMILIES AND FOR ALL FAMILIES IN THE UNITED STATES

	Participating Families			All Families	
NIT Support Level	Change in Hours[b]	Percent Change	Number of Families[c]	Change in Hours[b]	Percent Change
***NIT* Tax Rate 50 Percent**					
50 Percent of Poverty Level[a]					
Husbands	-104	-7.0		-4	-0.2
Wives	-92	-23.3		-2	-0.3
Total H/W	-196	-10.3	2.4	-6	-0.2
Female Heads	0	0.0	2.3	+16	+1.6
75 Percent of Poverty Level[a]					
Husbands	-106	-5.9		-19	-1.0
Wives	-110	-22.8		-19	-2.4
Total H/W	-216	-9.5	7.6	-38	-1.4
Female Heads	-47	-6.7	3.0	-23	-2 4
100 Percent of Poverty Level[a]					
Husbands	-119	-6.2		-47	-2.4
Wives	-130	-22 7		-50	-6.3
Total H/W	-249	-10.0	15.7	-97	-3.5
Female Heads	-99	-12.0	3.6	-69	-7.1
***NIT* Tax Rate 70 Percent**					
50 Percent of Poverty Level[a]					
Husbands	-136	-10.8		-2	-0.1
Wives	-111	-29.9		0	0.0
Total H/W	-247	-15 1	1.3	-2	-0.1
Female Heads	-10	-2.7	2.0	+20	+2.1
75 Percent of Poverty Level[a]					
Husbands	-157	-11.2		-9	-0.5
Wives	-126	-32.5		-5	-0.6
Total H/W	-283	-15 8	2.8	-14	-0.5
Female Heads	-47	-9.3	2.5	-12	-1.2
100 Percent of Poverty Level[a]					
Husbands	-164	-10.1		-23	-1 2
Wives	-144	-32.0		-18	-2.3
Total H/W	-308	-20.6	5.8	-41	-1.5
Female Heads	-95	-14.9	3.0	-52	-5.3

Notes. Average hours of work per year for all husbands in the United States before response = 1,999. Average hours of work per year for all wives in the United States before response = 793. Average hours of work per year for all female heads in the United States before response = 974. Total number of husband-wife families in the United States = 39.8 million. Total number of female-headed families in the United States = 4 9 million

[a]Poverty level was $5,000 per year for a family of four in 1974.

[b]Average change in hours of work per year due to *NIT*.

[c]Shown in millions.

do for participants), they decrease with the tax rate for both groups. This inverse relationship between the average *U.S.* response and the tax rate is an interesting and perhaps unexpected result that is a consequence of the fact that the number of participants decreases by an amount large enough to offset the effect of a larger response among participants. Thus, we find that the total disincentive effect of a nationwide *NIT* program is smaller under higher tax rate programs.

VI. Summary and Conclusions

Social experimentation is a relatively new research tool that is being used to assess the behavioral effects and costs of alternative public transfer programs. Its success as a research tool depends on developing an

empirical framework that exploits the advantages of experimental data (primarily, exogeneity of treatment) and at the same time accounts for the unique aspects of experimental design that create problems in the analysis (namely, nonrandom assignment, nonrepresentative samples, limited duration, the presence of other welfare programs, etc.).

In this paper, we present a framework for using experimental data to estimate the parameters of a labor supply response function. The nationwide aggregate labor supply effects of alternative *NIT* programs are obtained by applying these parameter estimates to a national data base. The results indicate that the labor supply responses to alternative nationwide *NIT* programs vary widely with the parameters of the program, and that for some programs, the aggregate labor supply responses are of considerable magnitude.

REFERENCES

H. J. Aaron, "Cautionary Notes on the Experiment," in Joseph Pechman and P. Michael Timpane, eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment*, Washington 1975, 88–114.

T. Amemiya, "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, Nov. 1973, *41*, 997–1016.

O. Ashenfelter and J. Heckman, "Estimating Labor Supply Functions," in Glen G. Cain and Harold W. Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973, 265–78.

——— **and** ———, "The Estimation of Income and Substitution Effects in a Model of Family Labor Supply," *Econometrica*, Jan. 1974, *42*, 73–85.

Glen G. Cain and Harold W. Watts, *Income Maintenance and Labor Supply*, Chicago 1973.

J. Conlisk and M. Kurz, "The Assignment Model of the Seattle and Denver Income Maintenance Experiments," res. memo. no. 15, Center Study Welfare Policy, SRI International, July 1972.

J. Dickinson, "Implicit and Explicit Preference Structures in Models of Labor Supply," disc. paper no. 331-75, Instit. Res. Poverty, Univ. Wisconsin, Dec. 1975.

———, "The Ashenfelter-Heckman Model and Parallel Preference Structures," disc. paper no. 411-77, Instit. Res. Poverty, Univ. Wisconsin, Dec. 1977.

Milton Friedman, *Capitalism and Freedom*, Chicago 1962.

R. Gronau, "Wage Comparisons—A Selectivity Bias," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1119–143.

R. E. Hall, "Wages, Income, and Hours of Work in the U.S. Labor Force," in Glen G. Cain and Harold W. Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973, 102–62.

———, "Effects of the Experimental Negative Income Tax on Labor Supply," in Joseph A. Pechman and P. Michael Timpane, eds., *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment*, Washington 1975, 115–47.

J. Heckman, "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

———, "Sample Selection Bias as a Specification Error (with an Application to the Estimation of Labor Supply Functions)," mimeo., Univ. Chicago, Apr. 1976.

——— **and Robert Willis,** "A Beta Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *J. Polit. Econ.*, Feb. 1977, *85*, 27–58.

E. D. Kalachek and F. Q. Raines, "Labor Supply of Lower Income Workers and the Negative Income Tax," in *Technical Studies*, The President's Commission on Income Maintenance Programs, Washington 1970, 159–86.

Michael C. Keeley, *The Economics of Labor Supply: A Critical Review*, forthcoming.

——— **and Philip K. Robins,** "The Design of Social Experiments: A Critique of the Conlisk-Watts Assignment Model," mimeo., Center Study Welfare Policy,

SRI International, Mar. 1978.

——— **et al.,** "The Labor Supply Effects and Costs of Alternative Negative Income Tax Programs," *J. Hum. Resources*, Winter 1978, *13*, 3–36.

David Kershaw and Jerilyn Fair, *The New Jersey Income Maintenance Experiment*, Vol. 1: *Operations, Surveys, and Administration*, New York 1976.

M. Kurz et al., "A Cross Sectional Estimation of Labor Supply for Families in Denver 1970," res. memo. no 24, Center Study Welfare Policy, SRI International, Nov. 1974.

——— **and R. G. Spiegelman,** "The Seattle Experiment: The Combined Effect of Income Maintenance and Manpower Investments," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 22–29.

——— **and** ———, "The Design of the Seattle and Denver Income Maintenance Experiments," res. memo. no. 18, Center Study Welfare Policy, SRI International, May 1972.

Robert J. Lampman, *Ends and Means of Reducing Income Poverty*, Chicago 1971.

M. Maxfield, "Estimating the Impact of Labor Supply Adjustments on Transfer Program Costs: A Microsimulation Methodology," mimeo., Mathematica Policy Res., Feb. 1977.

C. E. Metcalf, "Making Inferences from Controlled Income Maintenance Experiments," *Amer. Econ. Rev.*, June 1973, *63*, 478–83.

Jacob Mincer, *Schooling, Experience, and Earnings*, New York 1974.

G. H. Orcutt and A. G. Orcutt, "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," *Amer. Econ. Rev.*, Sept. 1968, *58*, 754–72.

Joseph A. Pechman and P. Michael Timpane, *Work Incentives and Income Guarantees: The New Jersey Negative Income Tax Experiment*, Washington 1975.

P. K. Robins and R. W. West, "Participation in the Seattle and Denver Income Maintenance Experiments, and Its Effect on Labor Supply," res. memo. no. 53, Center Study Welfare Policy, SRI International, Mar. 1978.

H. Ross, "A Proposal for a Demonstration of New Techniques in Income Maintenance," memo., Data Center Archives, Instit. Res. Poverty, Univ. Wisconsin, Dec. 1966.

R. G. Spiegelman and R. W. West, "Feasibility of a Social Experiment and Issues in Its Design," 1976 *Proc. Amer. Statist. Assn., Bus. and Econ. Statist. Sec.*, Washington 1976, 168–76.

J. Tobin, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, Jan. 1958, *26*, 24–36.

———, "Improving the Economic Status of the Negro," *J Daedalus*, Fall 1965, *94*, 878–98.

Harold W. Watts and Albert Rees, (1977a) *The New Jersey Income Maintenance Experiment*, Vol. II: *Labor Supply Responses*, New York 1977.

——— **and** ———, (1977b) *The New Jersey Income Maintenance Experiment*, Vol. III: *Expenditures, Health, and Social Behavior, and the Quality of the Evidence*, New York 1977.

U.S. Bureau of the Census, *Current Population Survey*, Mar. 1975 (data tape).

Public Utility Pricing under Risk: The Case of Self-Rationing

By JOHN C. PANZAR AND DAVID S. SIBLEY*

The existence of demand uncertainty poses important pricing and investment problems for planners of public enterprises. In particular, at given prices random surges in demand may exceed capacity. How then should price and capacity be set? The answer turns on how one assumes that capacity is rationed in the event of excess demand. Gardner Brown, Jr. and M. Bruce Johnson assumed that it could somehow be costlessly allocated among consumers on the basis of greatest willingness to pay. Under this assumption (and constant returns to scale) the welfare-maximizing price is simply equal to marginal operating cost, with the entire cost of capacity being born by the enterprise.

The Brown-Johnson article has generated several comments, criticisms, and extensions, beginning with that of Ralph Turvey. Subsequent authors[1] have tended to focus on the rather optimistic nature of their rationing assumption and the implications of the analysis for the financial viability of the firm. By modifying the assumptions about the way capacity is rationed, altering the analytical specification of the way uncertainty enters the demand function, and/or adding additional constraints to the problem, various authors have succeeded in obtaining "more palatable" solutions to the optimal pricing problem; that is, solutions in which the optimal price is greater than marginal operating cost.

An important shortcoming of this line of research is the failure to deal explicitly with the details of the rationing process itself. Since, in the event of excess demand, it is always assumed that available capacity *can* be utilized, it is clear that the analysts do *not* have in mind the disastrous consequences of a system failure, such as the New York City blackout of 1977, which may result when demand exceeds available capacity. In short, for the analysis to be relevant to the case of electric utilities, the literature has implicitly assumed that the utility is able to engage in some form of *load management* that enables it to avoid system failure.

Our approach in this paper is to combine optimal pricing decisions with a rudimentary form of load management, similar to some which have been in use for many years in Europe.[2] Each consumer subscribes to a particular level of capacity *before* the state of nature is revealed. He pays a capacity charge for the amount he subscribes to, and a usage charge for each unit actually consumed. If (and only if) his usage exceeds his subscribed capacity, a circuit breaker, or fuse, activates, curtailing further consumption. Consumers differ according to their willingness to pay for power; those with relatively high willingness to pay would, presumably, purchase relatively large fuse sizes. At the opposite extreme one could imagine consumers who do not find it worthwhile to purchase power at all.

The task of the utility in this setting is threefold: to set a usage price and a fuse price so as to maximize social welfare, while constructing enough capacity to meet the total required by consumers' choices of fuse size. By doing so, it can avoid any risk of system failure due to demand surges; load management is achieved by a decentralized system in which consumers ration themselves.

*Bell Laboratories and Council of Economic Advisors, respectively. The views expressed are our own and do not necessarily reflect those of the institutions with which we are affiliated. We would like to thank George Borts, Robert D. Willig, and an anonymous referee for helpful comments and suggestions.

[1]See, for example: Dennis Carlton; Michael Crew and Paul Kleindorfer (1978); Robert Meyer; Roger Sherman and Michael Visscher; Visscher.

[2]See Bridger Mitchell et al. for an interesting and thorough discussion of European pricing practices.

One might object that the usefulness of analyzing a rationing approach such as this depends crucially on whether or not the random variable entering demand is observable. For example, if it is temperature, which the utility and its customers can all observe easily, then the firm can do at least as well or better by just publishing a temperature-dependent price schedule giving for each temperature the price which clears the market at the extant total capacity. However, in order to do this the firm must be able to estimate demand accurately at each temperature, and the data needed to do this do not exist, at least in the United States. Considerable debate exists about price elasticities even for annual electricity demand (see Lester Taylor et al.). To be sure, pricing experiments are underway to refine our knowledge of electricity demand, but new data will not be forthcoming for some years. In any case, these experiments are not designed to yield temperature-related demand data sufficiently fine that one could compute temperature-related price schedules having any claim to reliability at the temperature extremes for which rationing would be in effect. As will be seen, the self-rationing approach we analyze requires very little information to implement and will even achieve a full optimum when all consumers react "similarly" to changes in temperature.

I. Consumer Behavior

Because our analysis focuses on the decisions of individual consumers, we must somehow "get behind" the aggregate stochastic demand curves used by previous authors. We assume that individual consumers act to maximize expected utility. There is a continuum of consumer types, each possessing a von Neumann-Morgenstern utility function $U(q, Y, t, \theta)$, where

q = quantity of power consumed
Y = income available for expenditure on other commodities
t = a random variable such as temperature
θ = index of consumer types.

The random variable temperature $t \epsilon [t_L, t_H]$ has continuous density $f(t)$; θ has positive density $g(\theta)$ for $\theta \epsilon [\theta_L, \theta_H]$.

The assumption that individuals maximize expected utility, although standard, raises a problem in specifying the objective of the planner. Previous authors, working with aggregate stochastic demand functions, have taken total expected consumers' surplus plus expected profits as their welfare measure. Such an objective is not necessarily compatible with expected utility maximization by individual consumers. One alternative approach would be to specify a social welfare function defined over the utilities of individual consumers and profits, and maximize its expectation. However, in order to make our analysis more comparable to the existing literature, we have chosen to impose a special form on $U(\cdot)$ which makes expected utility exactly equal to expected consumer's surplus. The analysis can then be carried out entirely in an expected surplus-maximizing format. Specifically, we assume that U takes the form

$$U(q, Y, t, \theta) = u(q, t, \theta) + Y$$

Without further loss of generality, we can rewrite $u(\cdot)$ as

$$u(q, t, \theta) = \int_0^q P(q', t, \theta) dq'$$

where $P(q, t, \theta)$ is the marginal willingness-to-pay function (or, the inverse demand function) of a consumer of type θ, giving the amount he would be willing to pay for an additional unit of the good in question. We assume that P is twice continuously differentiable and that

$$P_q \equiv \frac{\partial P}{\partial q} < 0, \; P_t \equiv \frac{\partial P}{\partial t} > 0, \; P_\theta \equiv \frac{\partial P}{\partial \theta} > 0$$

Thus, the ultimate goal of the consumer is to maximize his expected consumer's surplus under the following scenario: after the state of nature is realized the consumer is free to choose the amount of power he wishes to consume at a price p, so long as this does not exceed the size of the fuse (or circuit breaker) which he purchased beforehand at a price k per unit.

Key to our analysis is the *ex post desired demand function* giving the amount consumer θ would consume in state of nature t at price p in the absence of a self-rationing constraint. The *ex post* demand comes from maximizing consumer's surplus:

$$\max_{\{q \geq 0\}} \int_0^q P(q', t, \theta)dq' - pq$$

Necessary and sufficient conditions for a maximum at q^* are

$$(1) \quad P(q^*, t, \theta) - p \leqq 0, q^* \geqq 0, \quad q^* \cdot (P - p) = 0$$

so that individuals consuming positive quantities do so up to the point at which willingness to pay for a marginal unit just equals price. For $q^* > 0$, it is obvious that

$$(2) \quad \frac{\partial q^*}{\partial p} = \frac{1}{P_q} < 0, \frac{\partial q^*}{\partial t} = \frac{-P_t}{P_q} > 0, \quad \frac{\partial q^*}{\partial \theta} = \frac{-P_\theta}{P_q} > 0$$

Having derived the *ex post* desired demand function $q^*(p, t, \theta)$, we may compute expected surplus when the consumer is constrained. First, given any fuse size A, price p, and consumer type θ, there exists a state of nature $\hat{t}$ at which desired demand just equals the capacity of the fuse:

$$(3) \quad q^*(p, \hat{t}, \theta) = A$$

(equation (2) guarantees that $\hat{t}$ is unique). Thus, for $t \leqq \hat{t}$ consumption is given by the desired demand $q^*(p, t, \theta)$ and for $t > \hat{t}$, consumption is set equal to A. Expected surplus, then, is given by the expression

$$(4) \quad ES = \int_{t_L}^{\hat{t}} \left[\int_0^{q^*} Pdq - pq^*\right] f(t)dt + \int_{\hat{t}}^{t_H} \left[\int_0^A Pdq - pA\right] f(t)dt - kA$$

where, again, k is the unit price of fuse capacity. The first integral averages surplus in those states of nature in which rationing is not in effect ($q^* \leqq A$) and the second integral averages surplus over states in which the θth consumer is constrained by the size of the fuse (circuit breaker) which he selected *ex ante*.

The consumer, facing prices p and k, selects a fuse size A which maximizes his expected surplus. Using the implicit function $\hat{t}(p, A, \theta)$ defined by equation (3) the optimality conditions may be written

$$(5) \quad \frac{\partial ES}{\partial A} = \int_{\hat{t}}^{t_H} [P(A^*, t, \theta) - p] f(t)dt - k \leqq 0$$

$$A^* \geqq 0, \; A^* \cdot \frac{\partial ES}{\partial A}\bigg|_{A = A^*} = 0$$

where A^* is the optimal fuse size. For $A^* > 0$, equation (5) states that the consumer sets A^* at the level for which the price of an increase in fuse size k equals the expected excess of willingness to pay over the price of consumption p. For $A^* > 0$, equation (5) defines an implicit function $A^*(p, k, \theta)$ with partial derivatives given by

$$(6) \quad \frac{\partial A^*}{\partial p} = \frac{[1 - F(\hat{t})]}{\int_{\hat{t}}^{t_H} P_q dF(t)} < 0$$

$$\frac{\partial A^*}{\partial k} = \frac{1}{\int_{\hat{t}}^{t_H} P_q dF} < 0$$

$$\frac{\partial A^*}{\partial \theta} = -\frac{\int_{\hat{t}}^{t_H} P_\theta dF}{\int_{\hat{t}}^{t_H} P_q dF} > 0$$

where F is the cumulative of f.

An important implication of individual maximizing behavior is that all consumers choose to ration themselves in that, whenever k is positive, they select a fuse size which is strictly less than their greatest possible desired demand; i.e., $q^*(p, t_H, \theta) > A^*$. To see this, note that (5) must hold with equality for all consumers who choose to purchase a positive fuse capacity. If A^* were equal to $q^*(p, t_H, \theta)$, then, by definition, $\hat{t}(p, A^*, \theta) = t_H$. From (5), this would imply that $\partial ES/\partial A = -k < 0$, a contradiction. In other words, whenever fuses command a positive price, the consumer will not insure himself completely against the possibility of unsatisified desired demand.

II. Optimal Pricing

We study a planner whose objective is assumed to be to maximize the sum of (total) consumers' expected surplus plus expected profits via appropriate choices of p and k. Following the literature, we assume that power is produced at a constant cost b per kilowatt hour, and capacity (in kilowatts) costs β per unit installed. In our framework, the firm is committed to install capacity sufficient to cover the total of the maximum demands contracted for by each consumer, that is, productive capacity equals the *total* rated capacity of all fuses sold. Thus expected profits are given by

$$(7)\quad E\Pi = \int_{\hat{\theta}}^{\theta_H} \{[(p-b) \int_{t_L}^{\hat{t}(\theta)} q^*(p,t,\theta)f(t)dt + A^*\cdot[1-F(t)]] + (k-\beta)A^*(p,k,\theta)\}g(\theta)d\theta$$

where $\hat{\theta}$ denotes the marginal participating consumer type; that is, those who would choose a fuse with zero capacity. The term in square brackets gives the expected profit from usage of the θth consumer, with consumption given by desired demand $q^*(p,t,\theta)$ for $t \leqq \hat{t}(\theta)$ and by A^* for $t > \hat{t}(\theta)$. The second term inside the integral gives the net profit to the firm from consumer θ's choice of fuse (circuit breaker) size. It will be notationally convenient to express total expected profits as the sum of the expected profits resulting from serving each individual consumer. That is,

$$(8)\qquad E\Pi = \int_{\hat{\theta}}^{\theta_H} E\Pi(\theta)g(\theta)d\theta$$

where, obviously, $E\Pi(\theta)$ is just the integrand in (7).

The planner's decision problem can now be expressed as

$$(9)\quad \max_{\{p\geq 0, k\geq 0\}} EW \equiv \int_{\hat{\theta}}^{\theta_H} [ES^*(p,k,\theta) + E\Pi(p,k,\theta)]g(\theta)d\theta$$

where $ES^*(p,k,\theta)$ is given by equation (4) with $A = A^*$. Differentiating with respect to p and k we obtain the necessary conditions

$$(10)\quad \frac{\partial EW}{\partial p} = \int_{\hat{\theta}}^{\theta_H} \{\frac{\partial ES^*}{\partial p} + \int_{t_L}^{\hat{t}(\theta)} q^*(\theta)f(t)dt + A^*(\theta)[1-F(\hat{t})] + (p-b)\int_{t_L}^{\hat{t}(\theta)} \frac{\partial q^*}{\partial p} f(t)dt + \frac{\partial A^*}{\partial p}[(p-b)[1-F(\hat{t})] + (k-\beta)]\}g(\theta)d\theta - [ES^*(\hat{\theta}) + E\Pi(\hat{\theta})]g(\hat{\theta})\frac{\partial\hat{\theta}}{\partial p} \leq 0$$

$$p \geq 0, \qquad p\cdot\frac{\partial EW}{\partial p} = 0$$

$$(11)\quad \frac{\partial EW}{\partial k} = \int_{\hat{\theta}}^{\theta_H} \{\frac{\partial ES^*}{\partial k} + [(p-b)[1-F(\hat{t})] + (k-\beta)]\frac{\partial A^*}{\partial k} + A^*\}g(\theta)d\theta - [ES^*(\hat{\theta}) + E\Pi(\hat{\theta})]g(\hat{\theta})\frac{\partial\hat{\theta}}{\partial k} \leq 0$$

$$k \geq 0, \qquad k\cdot\frac{\partial EW}{\partial k} = 0$$

In writing (10) and (11) we have made extensive use of the definitional equation (3). We begin to simplify by noting that by definition of $\hat{\theta}$, $ES^*(\hat{\theta})$ and $E\Pi(\hat{\theta})$ are both equal to zero. (Obviously an individual who chooses to purchase a fuse of zero capacity neither receives a surplus nor makes any contribution to profits.) Noting from the definition of ES^* that

$$\frac{\partial ES^*}{\partial k} = \frac{\partial ES}{\partial k}\Big|_{A=A^*} = -A^*$$

$$\frac{\partial ES^*}{\partial p} = \frac{\partial ES}{\partial p}\Big|_{A=A^*} = -\int_{t_L}^{\hat{t}} q^* f(t)dt - A^*[1-F(\hat{t})],$$

we rewrite (10) and (11) as follows:

$$(12)\quad \frac{\partial EW}{\partial p} = \int_{\hat{\theta}}^{\theta_H} \{(p-b)\int_{t_L}^{\hat{t}(\theta)} \frac{\partial q^*}{\partial p} f(t)dt + \frac{\partial A^*}{\partial p}[(p-b)[1-F(\hat{t})] + (k-\beta)]\}g(\theta)d\theta \leq 0,$$

$$p \geq 0, \qquad p\cdot\frac{\partial EW}{\partial p} = 0$$

$$(13)\quad \frac{\partial EW}{\partial k} = \int_{\hat{t}}^{\theta_H} \{(p - b)[1 - F(\hat{t})] + (k - \beta)\} \frac{\partial A^*}{\partial k} g(\theta)d\theta \leq 0$$

$$k \geq 0, \qquad k \cdot \frac{\partial EW}{\partial k} = 0$$

Inspection of equations (12) and (13) reveals that these Kuhn-Tucker necessary conditions are satisfied by setting $p = b$ and $k = \beta$. To show that this is the only (p,k) pair which satisfies the necessary conditions is surprisingly somewhat complicated. That demonstration is relegated to the Appendix. Thus we have derived the intuitive result that the welfare-maximizing firm should charge marginal cost for both energy and fuses.

One way of looking at our approach comes from realizing that two complementary commodities are being consumed in our formulation of the problem: q itself, and insurance against being rationed. In this two-good model, if we look for an optimal price for each good, we should not be surprised by marginal cost prices. That is, clearly, what we get for the usage charge p. As for the "insurance" good, the cost to consumer θ of increasing his "coverage" by one unit is just β, the cost of increasing total capacity so as to increase $A^*(\theta)$. Visscher, for example, has only one price for the two "goods"; his usage charge for q reflects the rationing uncertainty imposed on consumers by his random and inverse rationing schemes, as well as the demand for q itself. That his optimal price is not a marginal cost price should not be surprising.[3]

III. Self-Rationing and Welfare

While the optimal prices in our self-rationing model have the intuitively appealing property of equaling the relevant marginal costs, we have not yet discussed how the maximized level of welfare compares to the maximum maximorum which could be attained if *ex post* rationing via greatest willingness to pay could in fact be accomplished *costlessly*. In general, our self-rationing scheme will result in a lower level of welfare. One obvious cause is that for a given realization of t some consumers may be constrained by their fuse while others still have some unutilized capacity. If available capacity could be costlessly allocated *ex post*, mutually advantageous "trades" could be effected in such a situation. Secondly, even when the realization of t is great enough so that all consumers are constrained by their fuses, there is nothing in our proposed solution that, in general, requires that marginal willingness to pay is equated across all consumers. If it is not, then we again have the potential for a welfare improvement if available capacity could be costlessly reallocated.

However, there does exist a broad class of consumer preferences under which the self-rationing approach *does* yield a full welfare optimum. This will occur whenever the marginal willingness to pay function is weakly separable in q and θ; i.e., when it can be expressed in the form

$$(14)\qquad P(q,\theta,t) = T[h(q,\theta),t]$$

The behavioral import of (14) is that consumers are similar in the way in which temperature changes affect their preferences.

In order to demonstrate our optimality result, we first show that all consumers become constrained by their fuses at precisely the same temperature when (14) holds. Equation (3) implicitly defines this critical temperature $\hat{t}(p,A,\theta)$. Use of the implicit function theorem and (2) readily establishes

$$(15)\qquad \frac{\partial \hat{t}}{\partial A} = -\frac{P_q}{P_t};\ \frac{\partial \hat{t}}{\partial \theta} = -\frac{P_\theta}{P_t}$$

When consumers select their fuse size optimally, A^* depends upon θ; thus we must look at the *total* derivatives of $\hat{t}$ with respect to θ:

$$\frac{d\hat{t}}{d\theta} = \frac{\partial \hat{t}}{\partial A} \cdot \frac{\partial A^*}{\partial \theta} + \frac{\partial \hat{t}}{\partial \theta}$$

[3]See M. G. Marchand for a "two-price" model of electric utility pricing under uncertainty which is similar to ours.

Upon substituting (6) and (15), this becomes

$$(16) \quad \frac{d\hat{t}}{d\theta} = \left(-\frac{P_q}{P_t}\right)\cdot\left(-\frac{\int_{\hat{t}}^{t_H} P_\theta dF}{\int_{\hat{t}}^{t_H} P_q dF}\right) - \frac{P_\theta}{P_t}$$

where all functions are evaluated at $q = A^*$. When P takes the separable form in (14), (16) becomes

$$\frac{d\hat{t}}{d\theta} = \left(-\frac{T_1 h_1}{T_2}\right)\cdot\left[-\frac{h_2 \int_{\hat{t}}^{t_H} T_1 dF}{h_1 \int_{\hat{t}}^{t_H} T_1 dF}\right] - \frac{T_1 h_2}{T_2} = 0$$

where the subscripts 1 and 2 refer to partial differentiation of h and T with respect to their first and second arguments.

Thus, we have met the first objection to the potential optimality of the self-rationing approach; there will be no temperature such that some consumers are being rationed while other have spare fuse capacity. The second objection was based on the familiar optimality requirement that any available quantity of a good must be allocated so that marginal willingness to pay is equal across all individuals receiving a positive allocation. In the self-rationing model, this condition is assured by equation (1) for all realized $t \leq \hat{t}$. (When no one is rationed, all consumers equate their marginal willingness to pay to price.) For realized temperatures greater than $\hat{t}$, we must examine the total derivative of $P(A^*(p, k, \theta), t, \theta)$ with respect to θ; i.e.,

$$(17) \quad \frac{dP}{d\theta} = P_q \cdot \frac{\partial A^*}{\partial \theta} + P_\theta$$

Using (6) and exactly the same steps and substitutions as above, it is easy to show that $dP/d\theta$ is equal to zero when (14) pertains.

It is also required that consumers who receive a zero allocation (in the self-rationing model, those who chose not to purchase a fuse) have an initial marginal willingness to pay no greater than the (equated) level of participating consumers. This condition is clearly satisfied by our assumption that P_θ is positive, since this implies that, for all $\theta < \hat{\theta}$ and $t\epsilon[t_L, t_H]$, $P(0, \theta, t) < P(0, \hat{\theta}, t)$, the level of the marginal willingness-to-pay function of the marginal participating consumer.[4]

Thus, it has been shown that, when consumers are "similar," the self-rationing approach leads to the fulfillment of those conditions which are necessary for any available level of capacity to be optimally allocated *ex post*. (We assume that said conditions are also sufficient.) All that remains to be shown is that the *level* of total capacity selected by consumers under the optimal self-rationing prices is in fact optimal. The easiest way to accomplish this is to demonstrate that capacity is the same as would be chosen in the simple Brown-Johnson (hereafter B-J) model, because if rationing according to greatest willingness to pay could indeed be costlessly accomplished, their approach would characterize the optimal capacity choice.

We proceed by setting $p = b$, since that is the optimal policy in both models. Let Z be the optimal level of capacity in the B-J model. Select that fuse price k_Z which results in Z being the *total* fuse capacity subscribed to by consumers.[5] Given weak separability, it has been demonstrated that the self-rationing scheme results in an optimal *ex post* allocation of *any* level of available capacity (including Z) *whatever* the actual temperature. Therefore, *expected* consumers' plus producer's surplus under a self-rationing scheme with $p = b$ and $k = k_Z$ must be exactly equal to the B-J optimum. (Analysis of the relevant first-order conditions would, of course, reveal that k_Z must equal β. Fortunately, we do not need to undertake that task to establish the result.)

[4]Since $A^*(p, k, \hat{\theta}) = 0$ by definition, we know from (17) that $P(0, \hat{\theta}, t) = P[A^*(p, k, \theta), \theta, t]$ for all $\theta \geq \hat{\theta}$.

[5]When $k \rightarrow \infty$, the size of total fuse subscriptions approaches zero. When $k = 0$, all consumers will subscribe to an arbitrarily large level of fuse capacity. Since from (6) we know that A^* is continuous in k, it must therefore, be possible to achieve any finite level of subscribed capacity (such as Z) via an appropriate choice of k.

IV. Implications and Conclusions

We have developed and analyzed a framework for optimal public utility pricing under risk which, while of basically the same structure as that in the literature, has a number of advantages over previous studies:

1) We solved for optimal prices p and k without resort to special (and restrictive) assumptions about the way uncertainty enters the demand function, whereas previous work has dealt exclusively with additive and multiplicative specifications. This may seem merely a technical nicety, but, as Dennis Carlton has recently demonstrated, the choice of specification can often make a substantial *qualitative* as well as quantitative difference in one's results.[6] Our analysis is valid for any (monotonic) specification of uncertainty.

2) Much of the criticism of Brown and Johnson's analysis has focused on the fact that, at their optimum, the firm makes *certain* losses equal to the entire cost of capacity. Under our scheme, the firm makes a *certain* profit of zero at the optimum; direct sales of q just cover operating costs, and revenues from sales of fuses just cover total capacity costs. In their original article, and (more specifically) in their response to Visscher, B-J suggested that the firm's losses could be recouped through the formation of a "futures market" for productive capacity. The present formulation can be viewed as accomplishing the solvency objective directly through self-rationing without appeal to "perfect risk or futures markets."

3) The problem of specifying the way available capacity will be rationed when demand exceeds capacity has caused great difficulties in the literature. Visscher rightly criticized Brown and Johnson for the unreasonableness of their assumption that available capacity could be costlessly rationed according to greatest willingness to pay. However, even Visscher's frameworks of random and inverse willingness-to-pay rationing may be too optimistic. In the case of electric power, for example, it is quite possible that the whole system may go down when demand exceeds capacity. The self-rationing framework avoids this difficulty. All that is required is that individuals understand the limitations imposed by their fuses.

4) The informational requirements of the present approach are minimal. Unlike earlier studies, where the firm must have complete information about consumer preferences in order to determine the optimal levels of capacity and (sometimes) price, in the self-rationing framework the firm needs to know only its own technological parameters. The self-interest seeking consumers do the rest.

We conclude with some suggestions for further research. It is clearly important to incorporate multiple periods into the analysis, since different users may have widely different time patterns of demand, and it would seem silly to allow an insomniac to "blow his fuse" at 4 A.M. It would also be desirable to analyze more flexible schemes, which might allow an individual to exceed his fuse by paying a surcharge, or allow the firm more discretion in the amount of capacity it actually installs. In addition, more sophisticated load management techniques may make relevant the analysis of additional innovative pricing schemes. Finally, an effort should be made to extend this kind of analysis to deal with the variable proportions productive technologies analyzed by Crew and Kleindorfer (1976), John Wenders, and Panzar in a peak load pricing context.[7]

[6]Carlton has shown that when the aggregate stochastic demand curve is multiplicatively separable in price and temperature, optimal pricing with random (inverse willingness to pay) rationing results in zero (positive) expected profits for the firm. Thus a shift from an additive to a multiplicative functional form eliminates the need for the second best analysis of Sherman and Visscher.

APPENDIX

Clearly k cannot be zero at the optimum, for with a zero fuse price all consumers

[7]See Robert Dansby and Derek McKay for analyses which attempt to incorporate these complications.

would choose a fuse at least as large as desired consumption at t_H. Thus $1 - F\{\hat{t}(\theta)\}$ would equal zero for all θ, and (13) would reduce to

$$(13') \quad \frac{\partial EW}{\partial k}\Big|_{k=0} = -\beta \int_{\hat{\theta}}^{\theta_H} \frac{\partial A^*}{\partial k} g(\theta)d\theta \le 0$$

which cannot be satisfied ($\partial A^*/\partial k \to -\infty$ as $k \to 0$). Since the optimal k must be positive, we can solve (13) and obtain

$$(A1) \quad (k - \beta) = -\frac{(p - b) \int_{\hat{\theta}}^{\theta_H} [1 - F(\hat{t})] \frac{\partial A^*}{\partial k} g(\theta)d\theta}{\int_{\hat{\theta}}^{\theta_H} \frac{\partial A^*}{\partial k} g(\theta)d\theta}$$

From (6), we note that

$$(A2) \quad \frac{\partial A^*}{\partial p} = [1 - F(\hat{t})] \frac{\partial A^*}{\partial k}$$

Substituting (A1) and (A2) into (12) yields

$$(A3) \quad \frac{\partial ES}{\partial p} = (p - b)\Big\{ \int_{\hat{\theta}}^{\theta_H} \int_{t_L}^{\hat{t}} \frac{\partial A^*}{\partial p} \cdot f(t)dt\, g(\theta)d\theta + \frac{H}{\int_{\hat{\theta}}^{\theta_H} \frac{\partial A^*}{\partial k} g(t)d\theta} \Big\} \le 0$$

$$p \ge 0, \qquad p \cdot \frac{\partial EW}{\partial p} = 0$$

where

$$(A4) \quad H = \Big[\int_{\hat{\theta}}^{\theta_H} [1 - F(\hat{t})]^2 \frac{\partial A^*}{\partial k} g(\theta)d\theta\Big] \cdot \Big[\int_{\hat{\theta}}^{\theta_H} \frac{\partial A^*}{\partial k} g(\theta)d\theta\Big] - \Big[\int_{\hat{\theta}}^{\theta_H} [1 - F(\hat{t})] \cdot \frac{\partial A^*}{\partial k} g(\theta)d\theta\Big]^2$$

which is nonnegative by the Cauchy-Schwartz inequality. Thus the bracketed term in (A3) is negative, and the inequalities can be satisified *only if* $p = b$. Substituting this result into (A1) yields $k = \beta$.

REFERENCES

G. Brown, Jr. and M. B. Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

D. Carlton, "Pricing with Stochastic Demand," *Amer. Econ. Rev.*, Dec. 1977, *67*, 1006–10.

M. A. Crew and P. R. Kleindorfer, "Peak Load Pricing with a Diverse Technology," *Bell J. Econ.*, Spring 1976, *7*, 207–31.

——— **and** ———, "Reliability and Public Utility Pricing," *Amer. Econ. Rev.*, Mar. 1978, *68*, 31–40.

R. E. Dansby, "Interruptible Service Options," unpublished paper, Bell Laboratories 1978.

M. G. Marchand, "Pricing Power Supplied on an Interruptible Basis," *Euro. Econ. Rev.*, July 1974, *5*, 263–74.

D. McKay, "Two Essays on the Economics of Electricity Supply," unpublished doctoral dissertation, California Instit. Technology 1977.

R. A. Meyer, "Monopoly Pricing and Capacity Choice Under Uncertainty," *Amer. Econ. Rev.*, June 1975, *65*, 326–37.

Bridger Mitchell et al., *Peak Load Pricing: European Lessons for U.S. Energy Policy*, Cambridge, Mass. 1978.

J. C. Panzar, "A Neoclassical Approach to Peak Load Pricing," *Bell J. Econ.*, Autumn 1976, *7*, 521–30.

R. Sherman and M. Visscher, "Second Best Pricing with Stochastic Demand," *Amer. Econ. Rev.*, Mar. 1978, *68*, 41–53.

Lester Taylor et al., *The Residential Demand for Energy*, Palo Alto 1977.

R. Turvey, "Public Utility Pricing and Output under Risk: Comment," *Amer. Econ. Rev.*, June 1970, *60*, 485–86.

M. Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand: Comment," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224–29.

J. T. Wenders, "Peak Load Pricing in the Electric Utility Industry," *Bell J. Econ.*, Spring 1976, *7*, 232–41.

A Generalized Model of Spatial Competition

By DENNIS R. CAPOZZA AND ROBERT VAN ORDER*

Recent research on the theory of the firm in economic space (for example, see M. L. Greenhut, M. Hwang, and H. Ohta, 1975; Capozza and Kazem Attaran) has argued that in a spatial context many of the conclusions of classical price theory are reversed. Since most firms operate in a spatial rather than a nonspatial environment, the result has far reaching implications. For example, Greenhut, Hwang, and Ohta have shown that some forms of (spatial) competition will result in higher prices than those realized under (spatial) monopoly. Capozza and Attaran have shown that spatial comparative statics can be different from nonspatial. In particular, cost increases (fixed, marginal, or transport) may cause *lower* market equilibrium prices.

Many of the spatial results appear to be sensitive to the assumption made concerning competition among firms (see Greenhut, Hwang, and Ohta, 1975; the authors, 1977a, b). The purpose of this paper is to investigate the extent to which spatial price theory replicates the results of classical price theory when there is free entry. To this end we analyze in detail two models of location under free entry which differ primarily in their assumption about price competition. An important aspect of the analysis is the use of the second model as a basis for analyzing monopolistic competition. Indeed, we believe that monopolistic competition can best be viewed as a result of the spatial dimension and that the results of monopolistic competition should be derived accordingly.

The traditional model of spatial competition follows the assumption in location theory (see, for example, Edwin Mills and Michael Lav; Martin Beckmann; Greenhut and Ohta; John Hartwick; Nicholas Stern among many others) that firms set price as if they are monopolists within their market area. In this paper a second model is developed from an assumption about pricing that is similar to that assumed by Harold Hotelling and followed by Arthur N. Smithies in the analysis of spatial duopoly. Under the Hotelling assumption each firm conjectures that all other firms will leave their price unchanged.[1] We believe that this assumption is closer to the way firms behave in most real world situations. A third assumption developed by Greenhut and Ohta is also analyzed.

To preview the conclusions, we find that many of the perverse results found in spatial price theory using the traditional assumption are also found in the second model using the Hotelling monopolistic competition assumption, but usually as special extreme cases. Thus while we are not able to rule out all perverse spatial results, we are able to show that the conditions under which they are likely to be realized are much more restrictive than earlier believed. The second model also leads to a generalized class of models encompassing a wide range of competitive reactions.

In the next section we consider the relationship between space and the theory of the firm. In addition we list the characteristics that we think reasonable spatial models should possess; these are subsequently used in discussing alternative models. The second section outlines the standard assumptions of location models and discusses the possible price reactions. The third and fourth sections analyze the two models. Finally, the last two sections present and evaluate a generalization of the earlier results.

*University of Southern California and Department of Housing and Urban Development, respectively. We are indebted to anonymous readers for helpful comments.

[1] The term "zero conjectural variation" is sometimes used to describe this assumption, see, for example, B. Curtis Eaton and Richard Lipsey.

I. Spatial Microeconomics

Before developing any spatial model, it is worthwhile to inquire into the relationship between spatial and nonspatial microeconomics. There are two essential distinguishing features of spatial competition. The first is transportation cost. If transportation were costless, firms would have no protection from spatially separated rivals. A firm 1,000 miles away would be just as formidable a competitor as one next door. As a result, space would be of no consequence; firms might as well all be located on the head of a pin, perfect competition would prevail. Transport cost gives the spatial firm its monopoly power over customers close to it.

The second essential feature is that average cost curves must be downward sloping over some range. The curve may have the negative slope, for instance, either because there are some fixed costs or because there are economies of scale in production. If average cost curves are nonnegatively sloped throughout the range, there will be no advantage to concentrating production at specific locations. Each consumer could produce his required consumption just as cheaply as any concentrated firm, and again spatial competition would disappear.

If either of the above features fails to hold in practice, there is no need for a spatial model. In fact, both are characteristic of most real world industries; but that is not the interesting point for model building. The issue that must be raised here is that nonspatial perfect competition would appear to be essentially a special case of imperfect spatial competition that arises if either of the above features is absent. Therefore, we should expect a reasonable spatial model to behave like the perfect competition model in the limit. In particular, we should expect the following characteristics:

1) As transport costs approach zero, perfect competition should be approached and price should approach marginal cost.

2) As fixed costs approach zero, concentrated production is less essential; spatial monopoly power is diminished; and again price should approach marginal cost.

The above would have to hold for a model to be valid over a wide range of values. In addition to 1) and 2), we might also expect that, at least for some values of the parameters (i.e., low transport and low fixed costs), the model should obey the intuitively appealing properties of nonspatial competitive theory that:

3) As costs (fixed, marginal, or transportation) rise, price should rise.

4) As demand density rises, firms should be able to take advantage of economies of scale. Price should fall in the long run.

5) As more firms enter the industry, there should be increased competition and price should fall.

The most widely used Löschian model of location violates all of the above. The inconsistencies of the Löschian model can be attributed to the fact that it is a special extreme case of a more general class of location models. In the next section we begin to develop the models.

II. The Assumptions of Location Models under Free Entry

Following the writings of August Lösch a somewhat standard set of assumptions has evolved for partial equilibrium models of location with free entry. We accept most of these, including:

ASSUMPTION 1: *There is a single commodity that can be produced in two-dimensional space with the same cost function (i.e., ubiquitous resources and technology).*

ASSUMPTION 2: *The cost function has constant marginal and fixed costs*

$$C = f + cX \tag{1}$$

where X = output; C = production cost; f = fixed cost; c = marginal cost.

ASSUMPTION 3: *Transport cost per mile is identical between any two points and equal to t units per mile.*

ASSUMPTION 4: *Potential consumers occupy a homogeneous unbounded plain at uniform density D.*

ASSUMPTION 5: *All consumers are identical and have a demand curve that is linear in delivered price* ($p + tu$).

$$x = a - b(p + tu) \tag{2}$$

where p = *the mill price;* x = *demand per consumer;* u = *distance to the firm; and* $a, b > 0$. *There is no price discrimination. Firms set the mill price, and transportation costs are paid by the consumer.*

ASSUMPTION 6: *Firms continue to enter until profits for all firms are driven to zero.*

To the above we add:

ASSUMPTION 7: *All market areas are circular.*

As is well known, circles will not cover a plane while hexagons will; however, the analysis of hexagonal market areas is not materially different. Extension to hexagonal areas is straightforward (see fn. 4). Assumption 7 narrows and simplifies the investigation. Since we are not concerned with the shape of market areas, there is no loss of generality by including this assumption as long as we restrict the discussion to two-dimensional market areas.

The final assumption, upon which our results depend, concerns the reaction of a firm to a change in a competitor's price. The usual assumption which has been analyzed extensively in this context is:

ASSUMPTION 8A: Löschian competition—*Each firm assumes its market area to be fixed, and sets prices like a monopolist within its market area.*

However, other assumptions are possible including:

ASSUMPTION 8B: Hotelling-Smithies (H-S) competition—*Each firm assumes the prices of competitors to be fixed.*

ASSUMPTION 8C: Greenhut-Ohta (G-O) competition—*Each firm assumes the price at the edge of the market area (border price) to be fixed.*

All three assumptions can be interpreted in terms of how the firm expects rivals to react to a change in its price. The differences are crucial to the results that follow and require some elaboration.

Note that the buyer has two decisions to make: where to buy and how much to buy. The first choice is made by buying where the delivered price (including transport costs) is smallest. If we consider two neighboring firms, the boundary of their market areas will be at a point where both firms' products sell for the same delivered price. Hence, if we have, as in Figure 1, two firms at a distance $\bar{U}$ apart, and draw each firm's delivered price (mill price plus transport) as a function of distance, then the boundary between firms will occur at the intersection of the two delivered price lines. At this intersection, consumers will be indifferent between the two sellers.

This intersection occurs at that radius R, satisfying

$$p + tR = \bar{p} + t(\bar{U} - R)$$

where p = the mill price of the first firm
$\bar{p}$ = the mill price of the second firm
$\bar{U}$ = the distance between firms
R = the radius of the first firm's market area

Solving for R we have

$$R = \frac{1}{2t}(\bar{p} - p + t\bar{U}) \tag{3}$$

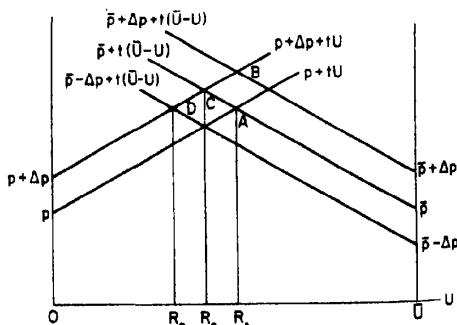



FIGURE 1. THREE PRICE REACTION ASSUMPTIONS

That is, the radius of the firm's market area depends on transport costs, distance to other firms, and its price relative to others' prices. Since we assume that all firms are identical, this formula applies in the same way to all its competitors, so that R is the "radius" of the market in all directions.

Consider Assumption 8. The representative firm located at $U = 0$ in Figure 1 is charging price p; its competitor at $U = \overline{U}$ charges price $\bar{p} = p$ initially. The market will be split between them, each having a market of radius $R_A = \overline{U}/2$. If the first firm should raise its price to $p + \Delta p$, then under the Löschian assumption the market radius must remain unchanged at R_A. However, the only way this result could be obtained would be if the rival raises its price at the same time to $\bar{p} + \Delta p$. Thus, one implication of the Löschian assumption is that the price reaction of a competitor is unitary, that is, $d\bar{p}/dp = 1$.

From (3) we have

$$\frac{dR}{dp} = \frac{1}{2t}\left(\frac{d\bar{p}}{dp} - 1\right) \tag{4}$$

and since $d\bar{p}/dp = 1$, we have $dR/dp = 0$, which is the Löschian case.

Under Assumption 8B of Hotelling and Smithies, the competing firms are assumed not to change their price ($d\bar{p}/dp = 0$). If so, the intersection of the price lines would move to point C in Figure 1 and the market radius of the first firm would fall to

$$R_C = \frac{1}{2t}(\bar{p} - p - \Delta p + t\overline{U}) < R_A$$

In this case, from (4) above, since $d\bar{p}/dp = 0$, we have $dR/dp = -1/2t$.

Under Greenhut-Ohta competition (8C) the border price is fixed.[2] In the context of our analysis the border price could remain constant when the first firm raises its price if rivals lower their price to $\bar{p} - \Delta p$ (i.e., $d\bar{p}/dp = -1$). The intersection would then be at point D in Figure 1. Market radius

[2]It should be noted that G-O competition was developed in the context of zonal pricing by competitors rather than by mill pricing. Thus the application in this context is somewhat strained. We retain the stylized version here because it helps to highlight the generalization in Section V.

TABLE 1—EXPECTED RADIUS AND PRICE RESPONSE

	Radius Change dR/dp	Conjectural Variation for Price $d\bar{p}/dp$
Löschian competition	0	1
H-S competition	$-1/2t$	0
G-O competition	$-1/t$	-1
Spatial Monopoly	$-1/t$	-

would fall to

$$R_D = \frac{1}{2t}(\bar{p} - p - 2\Delta p + t\overline{U}) < R_C$$

and since $d\bar{p}/dp = -1$, we have from (4), $d\overline{R}/dp = -1/t$ for the G-O case.

For completeness we consider also the spatial monopolist. The maximum price a consumer will pay is given by the price intercept of the demand curve, which is a/b. The monopolist's market will extend to the point where delivered price equals this maximum price or $p + tR = a/b$, whence $R = (1/t)((a/b) - p)$, and $dR/dp = -1/t$.

Thus each assumption can be converted to an equivalent assumption concerning either the change in market radius or the price reaction of competitor's price. Table 1 summarizes these cases.

The classification according to expected price reaction is particularly useful, because it offers comparison with the nonspatial market structure literature. In this light, Löschian competition is seen to be the spatial equivalent of noncompetitive oligopoly. Firms raise and lower price in unison either because of collusion or price leadership. H-S competition is comparable to monopolistic competition or competitive oligopoly.[3] Firms assume no reaction from

[3]The terminology employed in this paper is somewhat different from standard usage in the spatial economics literature. The Löschian model described here as spatial noncompetitive or collusive oligopoly is often referred to elsewhere as spatial monopolistic competition. We prefer to reserve the term "monopolistic competition" for the zero price reaction case to retain comparability with the nonspatial microeconomics literature. The term "spatial monopolistic competition" is used advisedly since many authors argue that monopolistic competition is not possible in a spatial context where there are only six immediate competitors.

competitors. G-O competition does not appear to have a nonspatial analog.

While it is clearly possible for any of these market structures to be realized in practice, our suspicion is that the most common market structure is something close to H-S competition. We have two reasons for favoring the H-S assumption: first, it incorporates the contest for market area which Löschian competition ignores by assuming market areas to be fixed; and secondly, it allows us to analyze spatial monopolistic competition, which many believe to be the most common industry structure. This analysis can be done in a relatively rigorous manner which gives a precise way of distinguishing individual consumer demand from demand faced by the firm. In addition, the model degenerates to perfect competition under intuitively appealing conditions. We turn now to the actual models.

III. The Löschian Model

Making use of Assumptions 1-8, the equilibrium price and radius can be derived. Total demand in a market of radius R will be[4]

$$(5)\quad X = D\int_0^R (a - b(p + tu))2\Pi\, u\, du = \Pi DR^2(a - bp - (2/3)btR)$$

Profits are

$$(6)\quad Y = pX - C$$

Substituting (1) and (5) into (6) we have

$$(7)\quad Y = \Pi DR^2(a - bp - (2/3)btR)\cdot(p - c) - f$$

From Assumption 6, profits will be zero in long-run equilibrium. This implies that

$$(8)\quad \Pi DR^2(a - bp - 2/3\, btR)\cdot(p - c) - f = 0$$

Equation (8) defines a curve in the (R,p)

FIGURE 2 THE LÖSCHIAN MODEL

plane for each level of fixed costs f. This curve is shown in Figure 2 for three levels of fixed cost and is designated the zero profit locus (ZPL). Two items are worth noting: first, this curve is invariant to the type of competition prevailing in the industry, so that it will be repeated in our discussion of H-S and G-O competition; second, the curve collapses around the point M as fixed costs rise. This is significant for three reasons: 1) M is the ZPL for $f = f_{max}$; for $f > f_{max}$ nonnegative profits are not possible, the ZPL is not defined, and no firms will survive. 2) The point M gives the price-radius combination a regional monopolist would choose[5] (i.e., the mill price at each production point and the optimal spacing of production points). 3) M also defines the

[4]Notice that equation (5) can be written $X = DA\bar{x}$ where A is market area and $\bar{x}$ is demand of a representative consumer. Written in this form, the equation is applicable to all market areas that are regular polygons. For example, if market areas are hexagonal $A = 3.45R^2$ and $\bar{x} = a - bp - .76tR$.

[5]Since the monopolist is unconstrained by competition, profits need not be zero. The monopolist's price is found by substituting $dR/dp = -1/t$ into equation (9) and solving for p which gives $p = 1/4\ (a/b + 3c)$. The monopolist's price is independent of density, fixed costs, and transport costs. It should be noted that this result holds only in a market of uniform density. See Dennis Heffley for a discussion of the spatial monopolist in a market with variable density.

maximum possible extent of the market given by $R_{max} = (3/4)t((a/b) - c)$. Radii beyond R_{max} are ruled out because beyond that point price is greater than $((a/b) - tR)$ and some customers would have negative demands.

A second condition is derived from the profit-maximizing behavior of firms. Maximizing (7) with respect to price gives

$$(9) \quad \frac{dY}{dp} = 0 = 2(p - c)(a - bp - btR) \cdot \frac{dR}{dp} - R(a - 2bp - (2/3)btR + bc)$$

Under Löschian competition $dR/dp = 0$ so that (9) becomes

$$(9') \qquad p = \frac{a}{2b} + \frac{c}{2} - \frac{tR}{3}$$

and gives the profit-maximizing price that a Löschian firm will charge in the short run given a market of radius R. It is shown as LL' in Figure 2.

Notice that the profit-maximization condition (9') and the zero profits conditions are satisfied twice for a given level of fixed costs. However, for demand to be nonnegative at the edge of the market area, it is required that $(a/b) - p \geq tR > 0$. Therefore all points to the right of the line $p = (a/b) - tR$ which is shown as ND (nonnegative demand) in Figure 2, are not admissible. If fixed costs are f_1, long-run equilibrium occurs at (R_1, p_1) where the profit-maximization line (LL') cuts the ZPL in the admissible region.

For markets with a radius larger than R_1, positive profits will be earned and firms will enter the industry. Entry will cause market areas to shrink and the industry will move towards long-run equilibrium. For radii smaller than R_1 profits are negative, firms will leave the industry, market radii will increase, and again the industry moves towards long-run equilibrium.[6]

It is straightforward to show that this model is inconsistent with the characteristics 1)–5) outlined above. We note the following behavior of the Löschian model.

1A) As transport costs approach zero, the profit-maximization line LL' becomes flat (note equation (9')), and price approaches the nonspatial monopoly price $((a/2b) + (c/2))$ rather than the perfectly competitive marginal cost price.

2A) As fixed costs approach zero, the ZPL expands, the equilibrium price rises and again approaches $p = (a/2b) + (c/2)$, the nonspatial monopoly price, rather than the marginal cost price.

3A) As fixed costs rise from f_1 to f_2 in Figure 2, the ZPL shrinks, and price falls in the long run to p_2. An increase in marginal costs shifts both the ZPL and the profit-maximization line (LL') so that the price

[6]Readers may be more familiar with the tangency solution that occurs in the (X,p) plane. The problem is actually three-dimensional in (X,p,R) space. The tangency solution arises graphically in the projection onto the (X,p) plane. Our graphical solution (Figure 2) is the projection onto the (R,p) plane. (See the authors, 1976.) The point (R_1, p_1) corresponds to the point of tangency.

TABLE 2—EFFECT ON PRICE IN THE LÖSCHIAN AND *SMC* MODELS

Parameter Change	Löschian Noncompetitive Spatial Oligopoly	*SMC* Normal $p < a/b - 4/3tR$	*SMC* Perverse $p > a/b - 4/3tR$	Spatial Monopolist
a	+	±	±	+
c	±	+	±	+
D	+	−	+	0
f	−	+	−	0
t	−	+	−	0

change is ambiguous. The comparative statics are developed more fully in Appendix A. The signs are summarized in Table 2.

4A) Appendix A shows that price increases when there is an increase in density. Intuitively this occurs because the higher density permits smaller market areas. The spatial demand curve is more inelastic for smaller market radii;[7] that is, the firm has more monopoly power and raises price.

5A) As firms enter, market area shrinks; the firm moves back along the profit-maximization line, and prices increase.

The above might be acceptable for a model of organized (price-leader) spatial oligopoly but is counter to what one might expect to hold in spatial monopolistic competition or unorganized oligopolistic competition. It is particularly important to note that in the limit as transport costs and fixed costs approach zero, the Löschian model approaches the nonspatial monopoly price rather than the perfectly competitive price. The model of the next section does approach perfect competition in the limit.

IV. A Model of Spatial Monopolistic Competition (*SMC*)

The model to be developed is one with Hotelling-Smithies competition (Assumption 8B) instead of Löschian competition. With this zero conjectural variation assumption we have shown above that $dR/dp = -1/2t$. Substituting in (9) gives the profit-maximization line for the *SMC* firm.[8]

FIGURE 3. A MODEL OF SPATIAL MONOPOLISTIC COMPETITION: NORMAL CASE

$$(9'') \quad p = c + \frac{tR(a - bp - 2/3\,btR)}{(a - bp)}$$

This can also be written as

$$(9''') \qquad p = c + \frac{tR\bar{x}}{x_o}$$

where $\bar{x}$ = average demand per person
x_0 = demand per person at $R = 0$

That is, profit-maximizing price is an additive markup over marginal cost where the markup is given by the second term on the right in (9‴).

Equation (9″) has an inverted U-shape in the admissible region and is shown in Figure 3. This profit-maximizing price curve starts at marginal cost ($p = c$, where $R = 0$) and rises to a maximum at $p = (a/b) - (4/3)tR$, then declines. This condition for a maximum is satisfied if average demand is unitary elastic with respect to transportation cost.[9] The *SMC* curve is downward

[7] The elasticity of the spatial demand curve in the Löschian model is

$$\eta = \frac{-bp}{a - bp - 2/3\,btR}$$

and becomes more inelastic as R decreases.

[8] To the individual firm R is endogenous. The firm's price equation is obtained from substituting (3) into (9″). In equilibrium if firms are identical they will move until they are equidistant. We can therefore interpret (9″) as an equilibrium condition with R set equal to one-half the distance between firms (assuming firms are not regional monopolists). For a brief discussion of short-run dynamics, see Appendix B.

[9] This elasticity is given by

$$\eta_{\bar{x},t} = \frac{d\bar{x}}{dt}\frac{t}{\bar{x}} = \frac{-\frac{2}{3}btR}{a - bp - \frac{2}{3}btR}$$

and is unitary if $p = a/b - 4/3\,tR$.

sloping, however, if average demand is elastic with respect to transport cost.

The zero profit locus (ZPL) is as before and is superimposed on Figure 3. Long-run equilibrium occurs at (R_3, p_3) if fixed costs are f_1. For market radii larger (smaller) than R_3 firms earn positive profits; new firms enter (exit) and move the industry towards equilibrium.

In this model two forces determine prices. First is the demand effect: the more elastic is average demand, the lower the price. The second is the competitive effect: the closer competing firms are, the greater is competition and the lower is price. When market radius is small (i.e., when firms are very close together) price is close to marginal cost. Competition forces price down to marginal cost in this model whereas in the Löschian model price approaches $(a/2b) + (c/2) > c$ as radius approaches zero.

There is, however, a point (given by $p = (a/b) - (4/3)tR$, the line KK' in Figure 3) where further increases in market area correspond to lower prices. For small market areas with densely packed firms, competition is important. However, as market areas approach the maximal size R_{max}, competition over the size of the market becomes less and less important since the consumer at the edge of the market is buying less and less (note that at R_{max} he is buying nothing and is not worth fighting over). Indeed, the SMC profit-maximization line approaches the Löschian line and intersects it at R_{max}. Hence, the Löschian-type effects noted above where monopoly power decreases as market area increases may tend to be important even in the SMC model. Thus, we cannot rule out a priori the "perverse" case in which entering firms raise rather than lower the equilibrium price. This perverse case occurs when firms are so spread out that individual firms are almost regional monopolists.

Mathematically we solve the model by substituting the profit-maximization equation (9″) into the zero profits equation (8) to obtain

$$(10)\quad (a - bp)^2(p - c)^3 = \left(\frac{t^2 f}{\Pi D}\right) \times \left[(a - bp - \left(\frac{2b}{3}\right)\left(\frac{t^2 f}{\Pi D}\right)\left(\frac{1}{(a - bp)(p - c)^2}\right)\right]$$

Explicit solution for price is difficult, but we can outline some properties of the long-run equilibrium. Note the following:

1) The zero profits locus is the same as that in the Löschian model (equation (8)).

2) Changes in f or D affect the zero-profits locus but not the profit-maximization curve (equations (8) and (9″)).

3) Since t, f, and D enter equation (10) as t^2f/D, solving for the impact of changes in one implies the effects of the others.

There are three possible cases:

CASE 1: *Degenerate case—nonnegative profits are not possible in the admissible region and no firms exist. This might occur if fixed costs were very large or population density very small.*

CASE 2: *Normal case ($p < (a/b) - (4/3)tR$, i.e., to the left of KK' in Figure 3)—zero-profits curve cuts the profit-maximization curve on the upward-sloping portion as in Figure 3. If additional firms enter, price falls.*

CASE 3: *Perverse case ($p > (a/b) - (4/3)tR$, i.e., to the right of KK' in Figure 3)—zero profits curve cuts the profit-maximization curve on the downward-sloping portion as in Figure 4. If additional firms enter, prices rise. This could occur for large fixed costs or small population density. However the implication is that transport costs are a large proportion of delivered price ($p + tR$) at the edge of the market area. Greenhut and Greenhut, Hwang, and Ohta (1976) present empirical evidence indicating that in general this is unusual. It would certainly be unusual in urban areas where the packing of firms is relatively dense.*

Only the latter two cases are of interest. In Table 2 the comparative statics have been summarized.

To derive the results consider first the effects of increasing fixed costs f. Rising fixed costs affect only the zero profits locus, shrinking it. Price rises if the profit-maximization line is upward sloping (normal

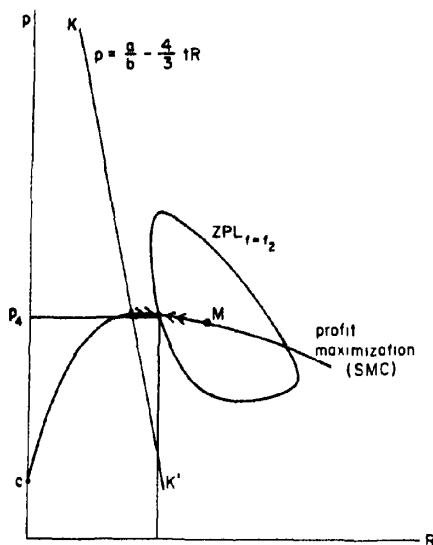



FIGURE 4. THE PERVERSE CASE OF THE *SMC*

case) and falls if the line is downward sloping. From property 3) the same holds for a rise in transport costs t, or a fall in density D. Increasing marginal cost c raises the profit-maximization line and shrinks the ZPL, raising price in the normal case, but with ambiguous effect in the perverse case. Similarly, a rise in the demand intercept a expands the zero profit locus and raises the profit-maximization line, leading to an ambiguous effect on price in the normal case and an increase in the perverse case.

The effects on price are quite different in the two cases. In the absence of a restriction on the sign of $(a - bp - 4/3btR)$ nothing can be said about the sign of the price changes. If we compare the signs for the *SMC* model with those for the Löschian model, we find that the Löschian signs are identical to the perverse case of the *SMC* model.

Note that characteristics 1)–5) hold for the *SMC* model.

1B) From (9″) as t approaches zero the profit-maximization line becomes horizontal at $p = c$, the marginal cost price.

2B) As fixed costs approach zero the intersection of the two curves moves back along the profit-maximization line and price approaches marginal cost.

3B, 4B) The comparative statics are normal if the intersection takes place on the rising portion of the profit-maximization line as discussed above. That is, increases in costs will raise price; increases in density will lower price.

5B) Entry forces price down if the profit-maximization line is upward sloping, i.e., in the normal case.

V. Spatial Competition—A Generalization

We have not developed a model with G-O competition, but it would be straightforward to do so. Since $dR/dp = -1/t$ under G-O competition, (9) would become

$$(9'''') \quad p = c + \frac{tR(a - bp - 2/3btR)}{2(a - bp - 1/2btR)}$$

This equation defines a curve in the (R,p) plane that starts at $p = c$ and rises to a maximum at M, the monopoly price, and falls afterward. The comparative statics in this model will be unambiguous (since the profit-maximization curve is monotonically increasing in the admissible region) and similar to the normal case of the *SMC*.

The three forms of competition suggest a generalization. Equation (9) defines a family of curves in the (R,p) plane which is shown in Figure 5. The two extreme cases in the

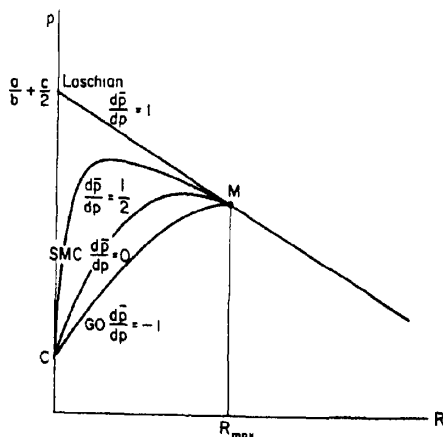



FIGURE 5. THE PROFIT-MAXIMIZATION CURVE UNDER DIFFERENT FORMS OF COMPETITION

figure are the Löschian profit-maximization curve where the price reaction $d\bar{p}/dp$ is plus one and the G-O curve where the reaction is minus one. Hotelling-Smithies competition is one intermediate case; but we can conceive of any number of other intermediate cases as the price reaction varies from 1 to -1. Notice that as the price reaction increases, the maximum point on the profit-maximization curve moves to the left. This implies that the perverse results noted for the Löschian model and for some values of the *SMC* model are more likely the larger the price reaction, that is, the greater the tendency for firms to collude.

A relevant question to raise at this point is which price reaction is most likely. We feel that in practice the value probably is fairly close to zero. In two-dimensional space, there are six competitors surrounding each firm. Each surrounding firm is a small proportion of the total competition faced by the firm. Therefore, it seems reasonable to argue that unless there is collusion, firms would not react dollar for dollar to price changes of just one of their competitors, and would probably have a coefficient much closer to zero than unity. In any case, the qualitative results of the *SMC* model would also hold for any value of price reaction between 1 and -1. We view the Löschian model as extreme, largely because it fails to meet the requirements set out earlier in the paper. In particular: it does not allow price to approach marginal cost as fixed or transportation costs approach zero. It unambiguously rules out the plausible cases where fixed and marginal cost increases raise price. It rules out the possibility that a density increase will decrease price. The argument that firms match price increases of rivals dollar for dollar seems implausible. As a model of collusive oligopoly, however, it is more acceptable.

The Greenhut-Ohta model has the advantage that it does give plausible and unambiguous comparative statics results, and handles the extreme cases properly. However, in the context of spatial competition with mill pricing, the implied negative price reaction seems unlikely. In another context (for example, under discriminatory or zonal pricing where a negative price reaction is not necessarily implied) it may be more appropriate.[10]

The Hotelling-Smithies competition does satisfy the criteria we have set up. While we should not expect price reactions to be exactly zero, we should (for reasons mentioned above) not expect it to be especially large, so that in the absence of collusion the zero conjectural variation assumption would appear to be a suitable approximation.

VI. Conclusion

We have shown that it is indeed the case that many perverse results are possible when the firm is placed in a spatial context. We have developed an imperfect competition model that shows that the perverse results appear under relatively "unusual" circumstances and that provides a convenient way of interpreting imperfect competition models as an outgrowth of spatial considerations.

In particular we have shown that the widely applied Löschian model, which exhibits the perverse results, can be interpreted as an extreme case of a class of models defined by the price reaction of firms. The perverse comparative statics results appear when the profit-maximization curve is negatively sloped in the (R,p) plane. The negative slope is more likely to be obtained if the price reaction $(d\bar{p}/dp)$ is close to one; if transport costs are a large proportion of delivered price or, more specifically, if average demand is elastic with respect to transportation cost; or if fixed costs are so large that firms are barely viable.[11]

All of these are clearly possible; but we have argued that in urban areas each would be an exception. Thus, the results of the theory of the nonspatial firm are usually

[10] See fn. 2 above.

[11] There is a fourth factor that influences the slope of the profit-maximization curve that has not been analyzed in this paper. Ohta has shown that the convexity of the individual demand curve also influences the slope. The negative slope is less likely the more convex individual demand is.

qualitatively similar to those for the spatial firm; however, there are conditions under which the traditional theory is violated.

We have not addressed the welfare issue raised by the lack of perfect competition. We have shown in our 1977a paper that total costs (production and transport) are not minimized. There are too many firms so that not enough advantage of scale economies is taken. This is a standard monopolistic competition result which we conjecture would carry over to the model here. We do not however, offer a proof.

Also, we have deliberately kept within a partial equilibrium framework. This framework is in the spirit of most work in monopolistic competition, but it is not entirely satisfactory. In particular, the distribution of population is assumed to be given; but one might expect the distribution to depend upon the price structure and transport costs; that is, if transport costs are high, consumers would concentrate at production sites and producers would cluster. We do not attempt to incorporate this aspect, viewing our analysis as a kind of market equilibrium curve, which should eventually be imbedded in a more general system.

APPENDIX A—THE COMPARATIVE STATICS OF LONG-RUN EQUILIBRIUM PRICES IN THE LÖSCHIAN MODEL

In the short run, with fixed market areas, price is given by equation (9') and is an increasing function of demand density a and marginal cost c but a decreasing function of transport cost t.

To determine the long-run price we combine equations (8) and (9') which gives the quartic in p

$$\text{(A1)}\quad \left(p - \frac{a}{2b} - \frac{c}{2}\right)^2 (p - c)^2 = \frac{ft^2}{9\Pi Db}$$

The solution is

$$\text{(A2)}\quad p = 4\frac{a}{b} + \frac{3c}{4} \pm \frac{1}{2}\left[\frac{1}{4}\left(\frac{a}{b} - c\right)^2 \pm \frac{4t}{3}\left(\frac{f}{\Pi Db}\right)^{1/2}\right]^{1/2}$$

The relevant root, after taking economic admissibility into consideration, is the positive one, so that

$$\text{(A3)}\quad p = \frac{1}{4}\left(\frac{a}{b} + 3c + 2Z\right)$$

where $Z = \left[\frac{1}{4}\left(\frac{a}{b} - c\right)^2 - \frac{4t}{3}\left(\frac{f}{\Pi Db}\right)^{1/2}\right]^{1/2}$

Differentiating with respect to the parameters we have

$$\text{(A4)}\quad \frac{\partial p}{\partial a} = \frac{1}{8bZ}\left(\frac{a}{b} - c + 6bZ\right) > 0$$

$$\text{(A5)}\quad \frac{\partial p}{\partial b} = -\frac{1}{8bZ}\left[2a\left(\frac{a}{b} - c + Z\right) - \frac{4}{3}t\left(\frac{f}{\Pi Db}\right)^{1/2}\right] \gtrless 0$$

$$\text{(A6)}\quad \frac{\partial p}{\partial c} = -\frac{1}{8Z}\left(\frac{a}{b} - c - 6Z\right) \gtrless 0$$

$$\text{(A7)}\quad \frac{\partial p}{\partial D} = \frac{t}{6DZ}\left(\frac{f}{\Pi Db}\right)^{1/2} > 0$$

$$\text{(A8)}\quad \frac{\partial p}{\partial f} = -\frac{t}{6DZ}\left(\frac{f}{\Pi Db}\right)^{1/2} < 0$$

$$\text{(A9)}\quad \frac{\partial p}{\partial t} = -\frac{1}{3Z}\left(\frac{f}{\Pi Db}\right)^{1/2} < 0$$

The sign of $\partial p/\partial a$ follows directly from the requirement that $a/b > p > c$ for firms not to shut down. The others follow from the requirement that Z be real for the solution to exist.

Intuitively the results come from the working of two effects. In Figure 2, changes in some parameters shift the short-run profit-maximization line LL', but at the same time the zero profits locus may expand or contract. An increase in a, which can be interpreted as an increase in income, shifts the profit-maximization line up and expands the zero profits locus. Thus we have the unambiguous positive sign. Increases in fixed costs shrink the ZPL and raise prices. Increases in transport cost lower the profit-maximization line and shrink the ZPL so that prices fall.

The sign of marginal costs is ambiguous because increases in c shift the profit-maximization line up, but shrink the ZPL. The resulting change in price depends on the slope of the profit-maximization line. If

transport costs are low, prices will rise; but if transport costs are sufficiently high, prices can fall.

Some of the results are somewhat counterintuitive. One would not expect prices to fall if fixed or marginal costs rose. Nor would one expect population density to have a positive effect and transport costs a negative effect. In general, prices are lower in urban areas than in low density rural areas because of higher volume and greater competition among sellers. Similarly, high transport costs are often cited as a cause of high prices not low prices, because of the lessened competition between firms.

Appendix B—Short-Run Dynamics

The "true" short-run price equation for the firm is (9″) with

$$R = \frac{1}{2}\overline{U} - \frac{1}{2t}(p - \overline{p})$$

substituted for R in the equation. This gives the firm's price as a function of the distance between firms (a proxy for the "number" of firms) and other firms' price. Short-run equilibrium is given in Figure 6.

The firm's price line is AA'. For prices to the left of p_e firms think they have charged more than their rivals price $\overline{p}$ but discover they have not and increase their estimates of $\overline{p}$, increasing prices. Similarly to the right of p_e prices must fall, and p_e is a stable equilibrium. It is possible to show that 1) if $\overline{p}$ is zero firms should set a positive price and 2) a unit increase in $\overline{p}$ leads to less than a unit increase in p so that AA' looks as it is drawn in the figure.

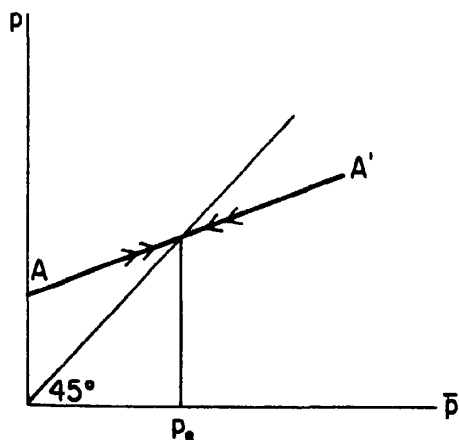



Figure 6. Price Adjustment in the Short Run

Then, in short-run equilibrium, p must equal $\overline{p}$ and R must equal $1/2\,\overline{U}$, so that (9″) can be intepreted as giving both profit maximization and (with $R = 1/2\,\overline{U}$) short-run equilibrium. We assume that short-run equilibrium is obtained "instantaneously" and suppress the dynamics of short-run adjustments. For a more complete discussion, see our 1976 paper.

REFERENCES

M. Beckmann, "Equilibrium vs. Optimum Market Areas," disc. paper no. 16, Brown Univ. 1970.

D. R. Capozza and K. Attaran, "Pricing and Spatial Dispersion of Firms Under Free Entry," *J. Reg. Sci.*, Aug. 1976, *16*, 167–82.

——— **and R. Van Order,** "An Equilibrium Model of Location and Pricing with Spatial Competition," work. paper no. 676, Univ. Southern California 1976.

——— **and** ———, (1977a) "A Simple Model of Spatial Pricing Under Free Entry," *Southern Econ. J.*, Oct. 1977, *44*, 361–67.

——— **and** ———, (1977b) "Pricing Under Spatial Competition and Spatial Monopoly," *Econometrica*, Sept. 1977, *43*, 1329-338.

Edward H. Chamberlain, *The Theory of Monopolistic Competition*, Cambridge, Mass. 1938.

B. C. Eaton and R. Lipsey, "The Principle of Minimum Differentiation Revisited," *Rev. Econ. Stud.*, Jan. 1975, *42*, 27–49.

Melvin L. Greenhut, *A Theory of the Firm in Economic Space*, Austin 1971.

——— **and H. Ohta,** "Spatial Configurations and Competitive Equilibrium," *Weltwirtsch. Achiv.*, Mar. 1973, *109*, 87–104.

——— **M. Hwang, and H. Ohta,** "Observations on the Shape and Relevance of the Spatial Demand Function," *Econometrica*, July 1975, *43*, 669–82.

———, ———, and ———, "An Empirical Evaluation of the Equilibrium Size and Shape of Market Areas," *Int. Econ. Rev.*, Feb. 1976, *17*, 172-90.

J. M. Hartwick, "Lösch's Theorem on Hexagonal Market Areas," *J. Reg. Sci.*, Aug. 1973, *13*, 213-22.

D. R. Heffley, "Transport Rate Changes and the Pricing Behavior of an Urban Spatial Monopolist," mimeo., Univ. Connecticut 1976.

E. M. Hoover, "Transport Costs and the Spacing of Central Places," 1970 *Papers Reg. Sci. Assn.*, Vol. 25, 255-74.

H. Hotelling, "Stability in Competition," *Econ. J.*, Mar. 1929, *39*, 41-57.

August Lösch, *The Economics of Location*, New York 1954.

E. S. Mills and M. R. Lav, "A Model of Market Areas with Free Entry," *J. Polit. Econ.*, June 1964, *72*, 278-88.

H. Ohta, "Spatial Competition, Concentration, and Welfare," memo., Aoyama Gakuin Univ. 1977.

A. W. Smithies, "Optimum Location in Spatial Competition," *J. Polit. Econ.*, Jan. 1941, *49*, 423-39.

N. H. Stern, "The Optimal Size of Market Areas," *J. Econ. Theory*, Apr. 1972, *4*, 154-73.

Fixed Wages, Layoffs, Unemployment Compensation, and Welfare

By H. M. POLEMARCHAKIS AND L. WEISS*

As tastes and technology change, the flexibility of wages and prices is a necessary condition for the competitive outcome to be efficient. This is apt not to be the case, however. In particular, since the time of John Maynard Keynes, economists have recognized that wages are relatively downward inflexible. Why this should be so is not transparent. At the most elementary level, this rigidity is accepted as an economic fact of life ultimately attributable to institutional constraints on the free movement of wages and prices: monopolies, labor unions, minimum wage laws, administered pricing policies, and the like. Rigid prices are assumed as an adequate description of modern economies; there is no need to look further. In the works of Robert Clower, Axel Leijonhufvud, and other disequilibrium theorists, it is argued that the slow diffusion of information about current opportunities allows prices to differ substantially from those which are market clearing.

More recently, Martin Baily (1974) and Costas Azariadis have suggested that real wages are stable over time as the necessary outcome of micro-economic optimizing behavior in a competitive labor market, even in a neoclassical environment which allows prices to instantly adjust to their market-clearing values. Workers, who are assumed to be risk averse, insure themselves against the possibility of future wage reductions by accepting a lower initial wage. Firms, by way of superior access to capital markets and more efficient sharing of risks, are assumed to be relatively more tolerant of risk, and hence can profitably supply such insurance.[1] This argument assumes that the opportunities for spreading the risk of wage reductions outside of the employment contract are limited; no insurance is available. Such contingent contracts need not, of course, be written or explicit. A firm acquires a reputation which affects its attractiveness to potential employees. This reputation includes not only the firm's history of wage reductions, but the firm's policy of layoffs, recalls, and overtime. The question that naturally arises is that of policy and welfare. Are such implicit contingent contracts socially desirable?

In this paper we employ a simple general equilibrium model to analyze the effects of alternative employment contracts. We demonstrate that the wage an employee receives in a given period need not correspond to his marginal product. Wages are not "correct" signals and there is no guarantee that labor will be allocated efficiently across industries. We consider the case of shifts in product demand and demonstrate that there is too little labor mobility and output response. The government, by subsidizing the costs incurred by individuals in changing jobs, can improve resource allocation. Since unemployment insurance is in effect such a subsidy, we conclude that it is desirable for the government to contribute at least partially to the payment of unemployment insurance benefits.

Our results hinge crucially on two assumptions. First, job changes are costly to the individual. These costs include the expenses involved in searching for a new job, costs of relocating, and perhaps the psychic costs of working in a new environment. Such costs are in addition to the expenses

*Department of economics, Harvard University; and Cowles Foundation for Research in Economics, Yale University, respectively. We wish to thank Kenneth Arrow, Martin Feldstein, Frank Hahn, Steven Shavell, Eytan Sheshinski, and Joseph Stiglitz for their criticism and suggestions.

[1]This argument requires that workers and firms have identical probability distributions concerning the states of nature.

borne by the firm in hiring and training the new worker. Thus for a firm to hire away a currently employed worker, it must offer a wage high enough to compensate for these additional expenses. The second assumption recognizes that firms are limited in their ability to discriminate between old and new workers, and generally must pay the former at least as much as the latter. Thus the firm must raise the wages of old workers as well, if it desires to expand employment by attracting people who already are employed. This confers an element of monopsony power to the firm in the short run when it seeks to expand in response to favorable demand shifts, even though over longer time periods the steady flow of new labor force participants allows the firm to take the supply of labor as perfectly elastic.

There is no unemployment in our model. Employees and employers both expect that, once terminated from a job, employees can instantly find a new job in the other industry at the prevailing wage, incurring a cost, denoted by c, in the process. These expectations are in fact fulfilled; in the jargon of the "new macroeconomics," they are "rational." However, a more realistic interpretation of the cost of transfer might reasonably include the time lost in locating a new job, the cost of "frictional unemployment." The substantive conclusions would not be affected.

We first analyze a situation where firms are prohibited from laying off employees. We demonstrate that a policy of flexible wages in one industry increases the desirability of wage flexibility in other industries. Thus, if firms are prohibited from dismissing employees and must rely solely on wage reductions to induce separations, two possible outcomes may exist. In one, wages are rigid and there is neither labor mobility nor output changes in response to shifts in demand. In the other, wages are flexible and labor flows to its most productive use. The fixed wage situation spreads risk efficiently, but allocates labor nonoptimally. In the flexible wage equilibria it is just the opposite. Both are full equilibria in the usual sense that a firm takes the prices of outputs, the wage agreements of the other firms, and the utility level it must offer to attract workers as data. Thus there is a tradeoff between the proper allocation of risk over time, and efficient allocation of resources at each instant of time.

If firms may offer contracts with the possibility of future terminations, then layoffs will be utilized exclusively to induce separations. This result, first suggested by Baily and others, merely requires employees to be risk averse to make a strategy of wage reductions unprofitable for the firm. The number of layoffs is less than what might be considered "socially" optimal. This suboptimality occurs due to the interdependence between the decisions of the two (or several) firms and the resulting divergence between the "social" and the "individual" criteria of optimality in firms' decisions. We demonstrate that unemployment insurance with less than complete experience rating lowers the cost of layoffs to the firm and encourages labor mobility. In the context of the model, a properly designed unemployment insurance program will yield a fully efficient allocation.

I. The Model

We consider a two-period world with uncertain second-period demand. There are two firms,[2] one of which manufactures good x, the other good y. The capital stock is fixed over both periods, and in each firm output equals the square root of labor input. Firms offer prospective employees contracts which specify a certain first-period wage and, contingent upon demand conditions, a second-period wage coupled with a probability of employment during the second period. A firm takes as given the wage contract offered by the other firm, the price of its output, and the utility level it must offer to employees.[3] It chooses an

[2] The number of firms in each industry is not important. For simplicity it is assumed to be 1.

[3] It is assumed here that the firm knows the workers' utility function as well as their expectations concerning the second-period states of nature. A weakening of this assumption would complicate the analysis without altering the results.

employment contract and labor demand to maximize expected profit. Uncertainty is generated by a random parameter in people's utility functions, which is symmetric between the two goods.[4]

Each person has a utility function equal in each period to $\tilde{\alpha} \log x + (1 - \tilde{\alpha}) \log y$.[5] Total utility is the sum of the utility levels of the two periods. During the first period, $\tilde{\alpha}$ is known to be 1/2. During the second period $\tilde{\alpha}$ will equal $(1/2 + \epsilon)$ or $(1/2 - \epsilon)$, each with probability of 1/2. For simplicity there is no transfer of wealth between periods. If during the second period an employee changes jobs, he incurs a monetary cost c. Each worker supplies a unit of labor inelastically each period, and there is one unit of aggregate labor.

We shall carry out the analysis from the point of view of one of the firms. We shall denote by p the first-period price of its output, by p_h the second-period price when demand for that output is high and by p_l when low. The corresponding prices for the other firm are, by symmetry p, p_l, and p_h, respectively. We shall let l denote the labor employed by the firm.

II. The Case with no Layoffs

We first analyze the case where the possibility of layoffs during the second period is not open to firms. Firms are, however, allowed to adjust wages to demand conditions. Why may a firm find it profitable to offer a wage contract with variable second period wages? Clearly, since workers are risk averse and the firm must provide them with a given utility level, wage variations increase expected costs. On the other hand, a firm may want to encourage separations via wage reductions in response to weak demand, and similarly, raise wages to attract additional workers when demand is strong. We observe that in order to affect voluntary labor mobility, wage rates must differ across industries by an amount at least as great as the cost of transfer, c.

It is clear that a firm's choice between variable and fixed second-period wages depends on the parameters of the production and utility functions—no general statement can be made. What is important to observe, however, is that a firm's decision is not independent of the choice of its competitor. Labor mobility depends on the difference between the wages offered by the two firms. As a result, the difference between high- and low-demand period wages that a firm must offer to induce labor mobility increases as this difference in the wage contract offered by its competitor decreases. It is this interdependence between firms' decisions that raises the possibility of multiple equilibria. Furthermore, there is a clear distinction between the individual rationality of an outcome, and its social rationality. Labor mobility may be unprofitable from the point of view of each firm individually, while desirable from the point of view of society as a whole, including the firms. Finally, the multiplicity of equilibria and the distinction between individual and social rationality do not depend on workers' risk aversion. As will be argued later, they can occur even in the case of risk-neutral labor.

We shall now consider explicitly the model presented in the previous section and demonstrate the points made above by choosing appropriate values for the parameters involved. First, let us suppose that both firms offer a wage of 1 in each period, independent of demand conditions, and know that second-period employment can neither be augmented or curtailed. Expected profits of each firm are given by $[p + 1/2(p_h + p_l)](l^{1/2} - 2l)$. The firm chooses input l to maximize expected profits, and so each firm demands labor equal to $(1/16)[p + 1/2(p_h + p_l)]^2$. By symmetry, each firm must demand half the aggregate labor supply. Equilibrium in the second-period goods' market requires

[4]This symmetry justifies the assumption of firms' risk neutrality, which is not otherwise apparent. Since one firm's high demand corresponds to the other firm's weak demand, the returns are perfectly and negatively correlated. Thus, an efficient capital market will value such returns as equal to their expectations.

[5]It is well known that such preferences imply an indirect utility function which is logarithmic in wealth, independent of prices. As a result, we may ignore portfolio-theoretic considerations from entering the worker's decision problem.

that $p_l/p_h = \alpha/(1-\alpha)$ where $\alpha = (1/2) - \epsilon$. Hence, for the labor and goods' markets to clear, $p = \sqrt{2}$, $p_h = 2\sqrt{2}\,(1-\alpha)$, $p_l = 2\sqrt{2}\,\alpha$, and each firm employs 1/2 units of labor to produce $1/\sqrt{2}$ units of output. In this situation, expected profits are 1 for each firm.

To prove that this is a Nash equilibrium, we must demonstrate that a firm, taking prices as well as the wage contract offered by its competitor as given, cannot increase its expected profits by offering an alternative wage contract. There are three possibilities open to the firm other than the fixed wage-constant employment contract, and they all involve variable second-period wages. The firm may want to increase labor employed during high demand, and decrease it during low demand; it may want to decrease labor employed when demand is low, but keep it at its first-period level when demand is high; finally it may want to increase labor employed when demand is high, but keep it at its first-period level when demand is low.

Let us consider the first alternative. The firms must offer a wage of $1 + c$ to attract workers under strong demand and $1 - c$ to induce separations under weak demand.[6] First-period wage w must be sufficient to guarantee employees the same level of utility as the other firm offers; i.e.,

$$\log w + 1/2 \log(1-c) + 1/2 \log(1+c) = \log 1 + \log 1 = 0$$

Hence $w = (1 - c^2)^{-1/2}$, which is greater than 1. Under this strategy, it is optimal to select labor demand in each period myopically. First-period employment is $(1/4)(p^2/w^2)$ or $(1 - c^2)/2$. Second-period employment is $2(1-\alpha)^2/(1+c)^2$ when demand is high and $2\alpha^2/(1-c)^2$ when demand is low. Expected profits are given by

$$\sqrt{1-c^2}/2 + (1-\alpha)^2/(1+c) + a^2/(1-c)$$

which must be less than the profits which accrue to the firm under fixed wages (i.e., 1) for the latter to be an equilibrium. That this is indeed possible can be seen by taking α to be .45 and c to be .15. For these values of the parameters, each firm can increase expected second-period profits by switching from a constant employment policy to a policy of upward as well as downward variable employment policy. However, the increased first-period labor cost associated with such a policy suppresses first-period profits by a greater amount, and hence renders variable wages unprofitable.

To complete the argument that a fixed wage and employment policy is a Nash equilibrium, we must show that for the same values of the parameters it is not profitable for the firms to follow a policy of increasing labor demand when demand is high, while maintaining the first-period employment level when demand is low. To follow such a policy, a firm must offer a wage w during the first period, w_h when demand is high, and $(1 - c)$ when demand is low. Since workers are not risk lovers, $w_h = w$; and since the contract must offer a utility of 0, $w = (1-c)^{-1/3}$, which is greater than 1. Low-demand labor demand will be chosen myopically. Labor demanded during the first period, which is also the labor employed during high demand, can be computed to be $[(2p + p_h)/(6_w)]^2$. At that level of employment, however, the marginal value product of labor during low demand is

$$p_l/2\sqrt{l} = \frac{2\sqrt{2}\alpha 6(1-c)^{-1/3}}{2[2p + p_h]} \approx .92$$

which is greater than $1 - c$ (or $= .85$). But then, for these values of the parameters (i.e., $\alpha = .45$, $c = .15$), the firm is going to maintain a constant level of employment, and hence it has no incentive to offer a contract with wage variability. By an analogous argument we can exclude the possibility of a contract involving variable wages and only increased employment during high demand. For the values $\alpha = .45$, $c = .15$ the constant wage employment contract is a Nash equilibrium.

[6]It is assumed that c is divided between the two goods in the same ratio as any other income. Alternatively c can be modelled as a utility cost without altering the analysis in any significant way.

Suppose now that each firm offers a wage of 1 in the initial period, and in the second period, $1 + c/2$ when demand is high and $1 - c/2$ when demand is low. Then, in the second period, each firm decides on labor demand after observing the price for its output. The high-demand firm chooses to employ $p_h^2/4(1 + c/2)^2$ and the low-demand firm $p_l^2/4(1 - c/2)^2$. Full employment requires that

$$\frac{p_h^2}{4\left(1 + \frac{c}{2}\right)^2} + \frac{p_l^2}{4\left(1 - \frac{c}{2}\right)^2}$$

be equal to 1. Equilibrium in the second-period goods market implies

$$\left(\frac{p_h}{p_l}\right)^2 = \left(\frac{1 - \alpha}{\alpha}\right)\left(\frac{1 + \frac{c}{2}}{1 - \frac{c}{2}}\right)$$

so that

$$p_h^2 = \frac{4(1 - \alpha)\left(1 - \frac{c^2}{4}\right)\left(1 + \frac{c}{2}\right)}{1 + c(\alpha - 1/2)}$$

and

$$p_l^2 = \frac{4\alpha\left(1 - \frac{c^2}{4}\right)\left(1 - \frac{c}{2}\right)}{1 + c(\alpha - 1/2)}$$

Firms' expected profits when offering a variable wage contract are

$$1/2(1 - c^2/4)(1/1 + c(\alpha - 1/2)$$

in the second period. The first period is identical to the previous situation of fixed wages; each firm earns 1/2 in period 1.

To demonstrate that this configuration is also a Nash equilibrium for the same values of the parameters, we must show that no alternative wage contract yields higher expected profits. The firm has three alternatives to consider. It may want to have a constant employment level, independent of demand conditions; it may want to increase employment in response to strong demand but keep it at its first-period level otherwise; finally, it may want to decrease employment in case of weak demand, but keep it at its first-period level otherwise.

Let us consider the first alternative. If a firm offered w in each period and maintained constant employment, it could earn $1/2w$ in the first period and $1/2(p_h + p_l)^2/8w$ in the second, where w yields the same utility as the variable wage; i.e.,

$$2\,log\,w = log\,1 + \frac{1}{2}\,log\left(1 + \frac{c}{2}\right) + \frac{1}{2}\,log\left(1 - \frac{c}{2}\right)$$

so that $w = (1 - c^2/4)^{1/4}$ which is less than 1. For the variable wage policy to be an equilibrium, it must yield higher expected profits for the firms than the policy of fixed wages; that is,

$$\frac{1}{2} + \frac{1}{2}\,\frac{1 - \frac{c^2}{4}}{1 + c\left(\alpha - \frac{1}{2}\right)}$$

must be greater than

$$\left(1 - \frac{c^2}{4}\right)^{-1/4}\left(\frac{1}{2} + \frac{\frac{1}{2}(p_h + p_l)^2}{8}\right)$$

For the values $\alpha = .45$ and $c = .15$, this is indeed the case.

The second alternative involves the firm's decreasing its second-period labor employment when demand is low, but maintaining an employment level equal to that of the first period when demand is high. The firms must offer a wage w during the first period, w_h when demand is high, and $(1 - c/2)$ when demand is low. Since workers are not risk lovers, w_h will be equal to w, which in turn must satisfy the equation

$$3/2\,log\,w = log\,1 + 1/2\,log\,(1 + c/2)$$

which means that $w = (1 + c/2)^{1/3}$. Optimal first employment is given by $[(2p + p_h)/6w]^2$. But then, during low demand, the marginal value product of labor is $p_l/2\sqrt{l}$ (approximately .909) which is less than $1 - c/2$ (or $= .925$). Hence, the firm will not find it profitable to decrease its labor force during low demand. As a result, it can do better by offering a fixed wage-fixed em-

ployment contract compared to the fully flexible employment policy. By an analogous argument, we can exclude the possibility of a contract involving only increased employment during high demand. For the values $\alpha = .45$, $c = .15$, the fully variable employment contract is a Nash equilibrium.

Although the example chosen requires employees to be risk averse, and this generally increases the range of parameters for which this phenomenon occurs, it is not by itself responsible. To see this, let us examine a situation in which both firms initially have fixed wages, and demand shifts are such that the price of the good in high demand exceeds that of the low-demand good by an amount between c and $2c$. Since, at a common level of employment, the marginal value product of a worker is proportional to the price of the output, it is possible to improve resource allocation (raise GNP) by transferring the marginal worker. But since his wage in the low-demand industry exceeds his marginal value product, he may not profitably be hired away, since at the prevailing prices his product in the high-demand industry is less than $1 + c$.

Having demonstrated the existence of two distinct equilibria, one would like to analyze how they compare from the point of view of the expected utility they provide. There are two aspects to be considered. One is whether, at equilibrium, the difference between the marginal value products of the workers in the two firms is equal to the cost of transfer, c.[7] The other is whether workers are insured against variability in the wage they receive. A situation of variable wages allocates labor so that the difference in productivity is exactly equal to the cost of transfer. This is easily observed since wages differ by c, and each firm is at its most preferred labor supply, where the wage is equal to the marginal product. However, this is achieved at the expense of exposing workers to risk, which could conceivably be insured against. It is clear that if the cost of transferring is small, the loss from not insuring (being of the order of c^2) is negligible compared to the costs of misallocation of labor.[8] If workers are risk neutral, no such ambiguity arises; variable wages are clearly superior. Furthermore, firms expected profits are higher in the variable wage regime.

III. Layoffs

Risk aversion on behalf of employees is sufficient to rule out wage reductions to induce separations. This is easily explained. Suppose a firm were to dismiss the same number of employees as wage reductions accomplish. The newly terminated employees would be no worse off and the remaining workers would not suffer a wage reduction. Since workers are risk averse, they would be willing to forego an amount in the first period greater than the expected gain in second-period income. Firms could offer the same level of utility and increase expected profits by insuring workers against the possibility of wage reductions.

Similarly, each firm will find labor costs lowest if it indemnifies workers against the costs of being terminated. Severance pay and supplementary unemployment benefits are common examples. Firms and workers both believe that workers can find another job at the prevailing wage after incurring the transfer expense c. The firms pay workers 1 in each period employed, and a severance pay of c if layed off. Because this policy exposes a worker to no risk, it is the cheapest way for a firm to offer the competitively determined level of utility. Each firm maximizes expected profits by choosing the appropriate labor input in the first period, and the number of workers terminated under conditions of weak demand. It takes the layoff policy of the other firm as given and is happy to employ the newly layed off when its own demand is strong, paying the common wage equal to one.

Depending on the magnitude of the shift in demand, two situations may exist. If the

[7] It is clear that the firm will never find it optimal to offer either strictly more than $1 + c$ or strictly less than $1 - c$ to attract or get rid of labor, respectively.

[8] Inefficient allocation of labor affects the worker's expected utility through higher variability of prices.

change in demand is great enough so that some workers are layed off, the marginal value product of a worker in the weak demand industry is equal to $1 - c$. Since it is optimal to pay a severance pay of c to workers who are terminated, the marginal cost of a worker who is employed by the firm in the first period is equal to the difference between what he receives if he is employed, 1, and what he receives if terminated, c, or $1 - c$. Since profit maximization implies that workers are hired in the initial period up to the point where expected marginal product equals the wage, the worker's marginal product under strong demand must be equal to $1 + c$ (his first-period wage 1 exactly equals his first-period marginal product). Thus, if attention is confined to symmetric equilibria, where each firm has the same employment policy, the difference in productivity between the marginal worker in the two industries is equal to $2c$. Since it costs only c to transfer the marginal worker, it is clear that resource allocation could be improved if the marginal worker were transferred from the weak-demand industry to the strong-demand industry. In this situation it is plain that it would not be profitable for the firm with strong demand to increase wages to attract new employees. It would have to offer at least $1 + c$ to encourage job transfers, at which point the new worker becomes a matter of indifference. However, it would have to raise the wages for all old workers to accomplish this. Risk aversion implies that such an uncertain rise in future wages is valued at less than its expected value, so that expected labor costs necessarily rise under this policy, rendering it unprofitable.

If the demand shifts are small, the competitive outcome may entail no layoffs or job transfers, even though some labor mobility would be desirable. Consider a change in tastes such that when output is maintained at first-period levels the marginal product of a worker in the low-demand industry is greater than $1 - c$, but less than $1 - c/2$. Since the expected marginal product of a worker must equal his wage, the marginal product of a worker in the high-demand industry is between $1 + c/2$ and $1 + c$. Thus, there is no incentive for the low-demand industry to terminate workers, because the marginal product of a worker exceeds his marginal cost $1 - c$. But the difference in productivity between the marginal worker in the two industries is greater than the cost of transfer, so welfare could be improved if some workers were transferred and suitably compensated.

Thus, the layoff equilibrium results in a less efficient allocation of labor than does variable wages. Workers, however, are insured against the possibility of wage reductions, so that there is an efficient allocation of risk. The layoff equilibrium is superior to the outcome under fixed wages, if some workers are in fact dismissed, as there is some labor mobility in the former, and in both there is an efficient allocation of risk.

IV. Unemployment Insurance

Several writers (Baily, 1974; Feldstein, 1976; Azariadis) have pointed out that the current poor method of experience rating implies a very large subsidy to layoffs. By experience rating it is meant that employers pay the actuarially fair value of the benefits accruing to their terminated employees, so that firms realize that they ultimately bear the costs of terminations when making layoff decisions. The current analysis suggests that such a subsidy may be desirable since it encourages labor mobility.

Compulsory full experience-rating insurance would have no effect in our model. Each firm would find it optimal to offer the benefits of such insurance on its own initiative, since any policy which exposes workers to risk is dominated by some certain income package. In the context of the model, however, if firms are required to pay only half the cost of such employment benefits, labor will be allocated in the most efficient manner. Firms will dismiss workers when demand is low up to the point where their marginal product is equal to their marginal cost $1 - c/2$, since $c/2$ must be paid in the form of higher unemployment insurance premiums if a worker is terminated. Since the marginal worker is hired at the point where his

expected product equals his wage, the product under favorable demand must equal $1 + c/2$. So the difference in productivity is exactly c, as efficiency requires. This is clearly a full optimum, since no worker is exposed to wage uncertainty. The insurance commission will make a loss, which must be covered by some means. This models the current situation where general tax revenue is sometimes used to finance unemployment benefits.

The 50 percent rule is meant to be illustrative. It rests crucially on the symmetry of production functions and the probability distribution of demand shifts between the two industries. In more complicated situations, there is no guarantee that a full optimum may be achieved by a policy which treats all firms identically. Nevertheless, the case for full experience rating is tenuous. In general, layoffs should be subsidized.

V. Conclusion

In the context of the model, the only competitive outcome involves firms offering fixed wage contracts with the possibility of layoffs. The outcome is suboptimal in that labor is misallocated between the two industries. If labor were not risk averse, this outcome would still be possible. However, it would also be possible in this case to have another equilibrium in which wages varied in response to demand. If this occurred, it would be efficient.

Our model suggests that wage flexibility in one industry compliments wage flexibility in other industries. Recently, Robert Hall has argued that the presence of a large nonentrepreneurial sector of the economy marked by rigid wages reduces the flexibility of wages in the residual competitive profit-maximizing sector. Our findings support this "spill-over of rigidity," as Hall terms it, at least in situations of unexpected demand changes between these two sectors (as opposed to the more macro-economic concept of shifts in the aggregate level of demand).

The downward inflexibility of wages is quite robust. This result, put forward by several earlier writers, merely requires that firms be less risk averse than their employees. The upward inflexibility of wages is an implication of our model. However, it is possible to conceive of more complicated situations where firms find it profitable to raise wages to attract additional workers. It is an interesting conjecture that this is responsible for an inflationary bias as the economy continually responds to demand shifts. However, such an implication is beyond the scope of the simple two-period nonmonetary model we have presented.

Perhaps the weakest ground upon which such analysis rests is the assumption of relative risk neutrality on behalf of firms. Two explanations for this have been advanced. The first argues that entrepreneurs are self-selected on the basis of their tolerance of (or actual preference for) risk. The other recognizes that the opportunities for diversification of risk are greater in the capital market than in the labor market. This argument is valid so long as the risks are not systematic, and unexpected changes do not affect all firms equally. It is these types of risk for which our model is relevant, such as shifts in preferences. For unexpected changes in the level of aggregate demand, which more or less affects all firms equally, only the first explanation is operative. The validity of this assumption is an empirical matter.

Although our model is quite specific, some conclusions appear to be of more general validity. Declining firms find it profitable to employ more workers than immediate considerations would imply. Expanding firms are frustrated in their quest for more labor by such forms of labor contracts whereby workers may receive more than their product. Thus a competitive economy might be less efficient in allocating labor during periods of fluctuating demand than during periods of relatively more stable demand.

REFERENCES

C. Azariadis, "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*,

Dec. 1975, *83*, 1183–202.

M. N. Baily, "Wages and Employment under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

———, "On the Theory of Layoffs and Unemployment," mimeo., Yale Univ. 1975.

R. Clower, "A Reconsideration of the Microeconomic Foundations of Monetary Theory," *Western Econ. J.*, Dec. 1967, *6*, 1–9.

M. Feldstein, "Unemployment Compensation: Adverse Incentives and Distributional Anomalies," *Nat. Tax. J.*, June 1974, *27*, 231–44.

———, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers*, Washington 1975, *3*, 725–45.

———, "Temporary Layoffs in the Theory of Unemployment," *J. Polit. Econ.*, Oct. 1976, *84*, 937–57.

R. Hall, "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers*, Washington 1975, *2*, 301–49.

John Maynard Keynes, *The General Theory of Employment, Interest, and Money*, New York 1936.

Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, London 1968.

Biased Screening and Discrimination in the Labor Market

By George J. Borjas and Matthew S. Goldberg*

The traditional economic analysis of discrimination is based on Gary Becker's study of taste discrimination by employers, employees, and consumers. More recent work by Kenneth Arrow (1972, 1973) has attempted to interpret intergroup wage differences in an alternative framework as a rational reaction to uncertainty in labor markets. His model of "statistical discrimination" demonstrates that when the screening process used to determine a worker's qualifications is costly, and prior expectations of productivity differ across race or sex groups, then wage differentials may arise between workers of identical productivity.

By implicitly assuming a perfect screening process, Arrow ignores a potentially important source of wage differentials, namely the fact that the screening process might be a more reliable predictor of productivity for one group than for another.[1] Our paper generalizes the Arrow model in two ways. First, in contrast to Arrow, we assume that all groups have *identical* distributions of productivity. Secondly, the screening process used by the firm to determine an applicant's productivity is "biased" in the sense that: a) members of various groups may "pass" the test in different proportions despite their identical productivity distributions; and b) the predictive power of the test might vary across groups. Our objective is to analyze the effects of these types of biases in the screening process on the wage differentials between different population groups.

I. The Model

Consider a perfectly competitive industry consisting of homogeneous firms. To help any particular firm determine the productivity of any given applicant, a screening process costing C dollars is undertaken.[2] As a result of the screening each worker is assigned a score: passing ($\hat{Q}$) or failing ($\hat{U}$). The firm is assumed to hire all those (and only those) individuals who pass the test, i.e., the $\hat{Q}$ applicants. Moreover, the population can be partitioned into two mutually exclusive productivity groups in terms of the qualifications necessary to perform the job in question: qualified individuals (Q) and unqualified individuals (U). Finally, the firm is assumed to know the distribution of productivity in each group. That is, the firm knows the probability, $P_i(Q)$, that an individual from group i is qualified for the job. For expositional simplicity we consider two race groups, whites ($i = w$) and blacks ($i = b$).

Within this framework, Arrow's model is obtained by making two specific assumptions. First, the test is a *perfect* predictor of productivity, hence $P_i(\hat{Q}) = P_i(Q)$.[3] Sec-

*University of California-Santa Barbara and University of Chicago, respectively. We are indebted to Ann P. Bartel, Barry R. Chiswick, Linda N. Edwards, Ira M. Goldberg, Samuel Schwarz, the managing editor of this *Review*, and an anonymous referee for useful comments and suggestions on previous drafts of this paper.

[1]In a rather different framework, Edmund Phelps has allowed for the reliability of the screening process to differ across groups. However, as Dennis Aigner and Glen Cain have shown, Phelps' assumption that *all* applicants are hired by the firm leads to the conclusion that expected wages are identical across groups, as long as the groups under consideration have the same expected productivity. Thus they argue that Phelps' model is inadequate as an explanation of discrimination in the labor market.

[2]Note that due to our assumptions of homogeneous firms the testing procedure must be identical across firms. See A. Michael Spence and Joseph Stiglitz for a more extensive discussion of the role of screening in the labor market.

[3]The assumption that the applicant's productivity is known with certainty upon testing is similar to the

ondly, the proportion of qualified whites is higher than the proportion of qualified blacks, $P_w(Q) > P_b(Q)$.[4] Given these assumptions, the following conditions must hold in equilibrium for a risk-neutral firm:

$$(1)\quad P_i(Q)[MP_i - w_i] = C \qquad (i = w, b)$$

where w_i is the competitive wage for group i, and MP_i is the value of marginal product of qualified group i workers. These conditions have a straightforward economic interpretation. Applicants who score $\hat{U}$ are not hired, and hence do not contribute to the gains of the firm. If, on the other hand, an applicant is predicted (correctly) to be qualified, the gain to the firm is given by the difference between the marginal product of a qualified worker and the wage. Weighting this difference by the probability of being qualified yields the expected return from screening one more applicant and, in equilibrium, this return must equal the marginal cost of screening. Assuming that *qualified* white workers and *qualified* black workers are perfect substitutes in production, $MP_w = MP_b$, the equilibrium wage differential is given by

$$(2)\quad w_w - w_b = \frac{C}{P_w(Q)P_b(Q)}[P_w(Q) - P_b(Q)]$$

If the proportion of qualified workers is larger in the white population than in the black population, qualified whites will receive higher wages than their equally qualified black counterparts. That is, the existence of uncertainty about productivity coupled with the costs incurred in determining that productivity will lead to firm behavior, which in effect makes qualified blacks "pay" for their group's smaller expected productivity.

Note that the wage differential vanishes when the two groups under consideration have the same productivity distribution. Our model (to be presented below) differs in that even abstracting from group productivity differences and setting $P_w(Q) = P_b(Q)$, we are able to generate wage differentials. This is accomplished by allowing for imperfect testing. In particular, we assume that screening processes are biased against blacks so that blacks having the same productivity as whites tend to perform worse on the test and/or the test is of a lower predictive power for blacks than for whites. This hypothesis has received extensive study in the psychological literature with respect to differences in IQ scores between whites and blacks. Two explanations have been advanced to explain this phenomenon. The first states that the score differential can be attributed to real differences in ability which might be due to genetic and/or environmental differences across races. The second argument states that the score differential is due to a "cultural bias" in the test: since intelligence tests are prepared by members of the "ruling" white middle class, it is inevitable that the test questions will be loaded in favor of experiences familiar to this group.[5] Because of our assumption of identical productivity distributions across races, it is this latter type of effect which we are considering.

The introduction of imperfect testing affects the equilibrium conditions described earlier since the firm must take account of the possibility that some individuals who pass the test will in fact be unqualified. Let MP_q (MP_u) denote the marginal product of a qualified (unqualified) worker. Due to our assumption that qualified (or unqualified) whites and blacks are perfect substitutes in production, both MP_q and MP_u are invariant to race. From the point of view of

common assumption in job search models that all job characteristics are known to the applicant upon searching the firm. See the authors for a relaxation of this assumption in job search models.

[4]Arrow introduced this model in terms of employer *beliefs* concerning the joint distribution of productivity and test scores in each racial group. Note, however, that it is inconsistent to have both perfect testing and beliefs which are erroneous; that is, beliefs that will not be confirmed by a perfect screening process. In order to be internally consistent, it must be assumed that employer beliefs concerning the productivity distribution in each race group are indeed justified.

[5]For a more extended discussion of these hypotheses, see Anne Anastasi and Arthur Jensen.

the firm, therefore, the expected marginal product of an employee from race group i is $\overline{MP_i} = P_i(Q \mid \hat{Q})MP_q + P_i(U \mid \hat{Q})MP_u$. This expected marginal product is a weighted average of the marginal products of qualified and unqualified workers. The weights sum to unity and consist of the probabilities that a worker is qualified (unqualified) given that he has passed the test. Thus the equilibrium conditions must be modified to

$$(3)\quad P_i(\hat{Q})[P_i(Q \mid \hat{Q})MP_q + P_i(U \mid \hat{Q})MP_u - w_i] = C \qquad (i = w, b)$$

From equation (3) we obtain the market wage differential:

$$(4)\quad w_w - w_b = [P_w(Q \mid \hat{Q}) - P_b(Q \mid \hat{Q})] \cdot (MP_q - MP_u) + \frac{C}{P_w(\hat{Q})P_b(\hat{Q})} \cdot [P_w(\hat{Q}) - P_b(\hat{Q})]$$

If the bias is such that a black applicant has a smaller probability of passing the test, despite the absence of group productivity differences, then the second term in (4) will be positive. This term represents the "cost effect" of biased testing, and it will favor whites. It is important to note the similarity between the cost term here and the wage differential in the Arrow model as given by equation (2). Either imperfect testing or true differences in the productivity distributions will generate a term of this form, and thus the two models yield identical predictions with respect to the cost effect of discrimination.

Only imperfect testing, however, creates the additional "productivity effect" given by the first term in equation (4). The sign of this term will depend upon how $P_i(Q \mid \hat{Q})$ differs across races and, of course, these conditional probabilities are related to the joint distribution of productivity and test scores. To examine how the screening bias affects this joint distribution, a change of variables is useful. In particular, define a random variable Y_i which is set equal to unity if the individual is truly qualified and zero otherwise. Similarly, let Z_i equal unity if the individual passes the test and zero otherwise. Given these definitions, we can then measure the predictive power of the test by computing the correlation coefficient between the random variables Y_i and Z_i yielding:[6]

$$(5)\quad r_i = \frac{(P_i(\hat{Q}))^{1/2}}{(P_i(\hat{U}))^{1/2}} \frac{(P_i(Q|\hat{Q}) - P(Q))}{(P(Q)P(U))^{1/2}} \qquad (i = w, b)$$

Given equation (5) we can solve for $P_i(Q \mid \hat{Q})$ and substitute into the wage differential in (4) yielding:

$$(6)\quad w_w - w_b = (P(Q)P(U))^{1/2} \cdot \left\{ r_w \frac{(P_w(\hat{U}))^{1/2}}{(P_w(\hat{Q}))^{1/2}} - r_b \frac{(P_b(\hat{U}))^{1/2}}{(P_b(\hat{Q}))^{1/2}} \right\} \cdot (MP_q - MP_u) + \frac{C}{P_w(\hat{Q})P_b(\hat{Q})} \cdot [P_w(\hat{Q}) - P_b(\hat{Q})]$$

Therefore the productivity effect of biased screening is seen to depend upon the racial differences in r_i and $P_i(\hat{Q})$. To isolate the separate effects of these two variables it is illuminating to first consider two special cases.

CASE 1: *Suppose that the screening process has the property that* $r_w > r_b$ *and* $P_w(\hat{Q}) = P_b(\hat{Q})$. *Thus while whites and blacks pass the screening process with equal probability, the test performs its task of "matching" qualified applicants* $[Y_i = 1]$ *with passing scores* $[Z_i = 1]$ *more reliably for whites than for blacks.*

A simple example will illustrate this point. Consider a firm which screens four applicants from each race group, and suppose that the distributions of Y_i (productivity) and Z_i (test scores) are as given in Table 1. It is clear that although $P_w(\hat{Q}) = P_b(\hat{Q}) = .5$, the test does a much better job

[6]It can easily be shown that Y_i and Z_i are characterized by the following properties: $E(Y_i) = P(Q)$, $E(Z_i) = P_i(\hat{Q})$, $var(Y_i) = P(Q)P(U)$, $var(Z_i) = P_i(\hat{Q})\,P_i(\hat{U})$, $cov(Y_i, Z_i) = P_i(Q \cap \hat{Q}) - P(Q) \cdot P_i(\hat{Q})$. Note that we omit the subscript i from the probabilities $P(Q)$ and $P(U)$ since the productivity distribution is invariant to race.

TABLE 1

Whites		Blacks	
Y_w	Z_w	Y_b	Z_b
1	1	1	1
1	1	1	0
0	0	0	1
0	0	0	0

of predicting the productivity of white applicants. In fact, the test predicts perfectly for whites [$r_w = 1$], yet the predictions for blacks are totally random [$r_b = 0$]. If the screening process has these properties, it can be seen from (6) that the productivity effect is positive and the cost effect vanishes. Since the white applicants hired are likely to be of better quality, white workers in the firm will have a higher expected marginal product, $\overline{MP}_w > \overline{MP}_b$, and hence a higher wage.

CASE 2: *Suppose the screening process is such that $P_w(\hat{Q}) > P_b(\hat{Q})$ and $r_w = r_b$. Thus although a greater proportion of white applicants receive passing scores, the ability of the test to predict qualifications is equal for both groups.*

The plausibility of this case is illustrated by the example in Table 2. We find that although $P_w(\hat{Q}) > P_b(\hat{Q})$, the correlation coefficients are equal, $r_w = r_b = .58$. Since two of the four applicants from each race are truly qualified [$Y_i = 1$], a perfect testing procedure would grant passing scores to precisely these individuals. Due to the bias, however, one of the two qualified blacks is erroneously assigned a failing score, while one of the two *unqualified* whites is assigned a passing score. Thus the test is too selective for blacks, dilutes the quality of the white labor force, and these errors (although opposite in nature) have identical effects on the correlation coefficients.

TABLE 2

Whites		Blacks	
Y_w	Z_w	Y_b	Z_b
1	1	1	1
1	1	1	0
0	1	0	0
0	0	0	0

Given these properties, we see from equation (6) that the first term is negative, thus this type of biased testing produces a productivity effect which *favors* blacks. The reason is that by being more selective for black applicants, those blacks who do pass the test are more likely to be qualified than their white counterparts in the firm, hence $\overline{MP}_b > \overline{MP}_w$. Therefore, abstracting from the cost effect, we find that a testing bias in which firms are much more selective in screening blacks actually *improves* the relative position of black workers hired.[7]

CASE 3: *In general, a biased test will, of course, create a cost effect favoring whites, as well as a productivity effect ambiguous in sign. In fact, the productivity effect will be positive, zero, or negative depending on*

$$\frac{r_w^2}{r_b^2} \gtreqless \frac{P_w(\hat{Q})\,P_b(\hat{U})}{P_b(\hat{Q})\,P_w(\hat{U})} \tag{7}$$

We have seen that there are two opposing influences on the relative productivity of black workers. First, by being more selective in the hiring of blacks, those blacks who are hired are likely to be of better average quality. However, this effect is counterbalanced by the fact that the scores of black applicants may be less informative than those of whites, so that being selective and hiring only the highest ranked blacks need not necessarily improve the expected productivity of black workers.

[7]It can be shown that the productivity effect will exist only if $0 < r < 1$, where r is the common level of the correlation coefficient. If the screening process provides no useful information on the applicant's productivity, then the random variables Q and $\hat{Q}$ are statistically independent and $P_w(Q \mid \hat{Q}) = P_b(Q \mid \hat{Q}) = P(Q)$. No productivity effect exists since all workers hired, whether white or black, are randomly chosen by the firm. Hence biased screening must worsen the relative position of blacks through the cost effect. Similarly, if the test were of perfect quality all those individuals who passed, white or black, would be qualified with certainty. Again, the productivity effect would vanish.

It is important to emphasize that the screening biases discussed in this paper change the expected productivity of workers from each race group *within* the firm despite the fact that the population distributions of productivity are identical. These differences in productivity will in turn affect the interpretations of observed wage differences between black and white workers. In particular, suppose that the net effect of biased screening is that $\overline{MP}_b > \overline{MP}_w$ (as in Case 2 above). An important empirical implication of this result is that the observed market wage differential *underestimates* the true extent of discrimination. That is, since biased screening leads to the hiring of superior blacks, in order to compute the true magnitude of discrimination we must take the observed wage differential and add onto it the difference in expected productivity between races. Alternatively, if $\overline{MP}_b < \overline{MP}_w$, the market wage differential would *overestimate* the true extent of discrimination.

II. Summary

Statistical discrimination models have provided an explanation of why information on race is rationally taken into account by profit-maximizing employers. We have expanded the analysis by considering the case in which the firm uses a screening process which does not provide perfect information on an applicant's productivity and which is biased against members of a particular race group. We considered the two consequences of this bias on the screening process: First, the bias might result in one race group (for concreteness, blacks) obtaining lower scores despite the fact that the productivity distribution is invariant to race. Secondly, the bias might also affect the quality of the test in the sense that black scores would be less reliable measures of productivity. It was shown that by introducing the realistic concept of screening bias, wage differentials between black and white workers could be explained without recourse to assumptions of differential ability distributions across groups.

REFERENCES

D. J. Aigner and G. G. Cain, "Statistical Theories of Discrimination in Labor Markets," *Ind. Labor Relat. Rev.*, Jan. 1977, *30*, 175–87.

Anne Anastasi, *Psychological Testing*, New York 1968.

K. J. Arrow, "Some Mathematical Models of Race Discrimination in the Labor Market," in Anthony H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972.

——, "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton 1973.

Gary S. Becker, *The Economics of Discrimination*, Chicago 1957.

G. J. Borjas and M. S. Goldberg, "The Economics of Job Search: A Comment," *Econ. Inquiry*, Jan. 1978, *16*, 119–25.

Arthur R. Jensen, *Educability and Group Differences*, New York 1973.

E. S. Phelps, "The Statistical Theory of Racism and Sexism," *Amer. Econ. Rev.*, Sept. 1972, *62*, 659–61.

A. Michael Spence, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*, Cambridge 1974.

J. E. Stiglitz, "The Theory of 'Screening,' Education, and the Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

Derived Demand and Distributive Shares in a Multifactor Multisector Model

By DAVID BIGMAN*

Most studies on the derived demand for factor inputs at the aggregate, nationwide level, and on the behavior of distributive shares in national income, form an analytical framework at the aggregate level which is parallel with the one at the industry level, by defining an aggregate production function for the economy as a whole. Unfortunately, the conditions for intersectoral aggregation are extremely restrictive and such an aggregate production function will be consistent with the sectoral technologies if and only if factor intensities are identical across sectors.[1] For most practical purposes, these conditions must be considered inappropriate.

In his skeptical note on this approach, Robert Solow remarked that

> ... [this] theory is in the first instance a microeconomic one. Between production functions and factor-ratios on the one hand, and aggregate distributive shares on the other lies a whole string of intermediate variables: elasticities of substitution, commodity-demand and factor-supply conditions, markets of different degrees of competitiveness and monopoly, far-from-neutral taxes. It is hard to believe that the theory offers any grip at all on the variability of relative shares as the data changes—in fact this may be viewed by some as a symptom of its emptiness. [p. 620]

A more comprehensive approach would require carrying out the analysis within the framework of a multi-sector model in which shifts in factor prices or factor supplies affect all the sectors simultaneously, and to examine the effects on the aggregates via the equilibrium conditions in the product and the factor markets. This approach is not constrained by the restrictive conditions on intersectoral aggregation of the functional relationships between inputs and outputs and permits the incorporation of demand and supply conditions in the various markets.

The purpose of this paper is to take a first step in that direction by suggesting the elements of the theory of derived demand and distributive shares within the generalized multifactor multisector model. The questions considered in the paper are the effects on the aggregate quantity demanded of a factor and on its absolute and relative share in national income of a change in its own price or in other factor prices.

As is well known, in a single-sector model the trend in the relative share of a given input depends on the distribution of the corresponding elasticity of substitution on either side of unity.[2] In the multisector framework the answer is substantially different and involves also the price elasticities of the demands for the products. Our results dispel much of the myth surrounding the Cobb-Douglas production function by indicating that even when in each sector of the economy the production process can be characterized by this functional form, this is neither a necessary nor sufficient condition for the constancy of the aggregate income shares.

I. The Model

The economy in the model consists of n sectors. To simplify the exposition interindustry relationships via the flow of intermediate inputs are disregarded and it is

*Department of political economy, Johns Hopkins University. I am indebted to Carl Christ, Charles Hulten, and an anonymous referee for helpful comments that have led to improvements in substance and exposition.

[1]See H. A. J. Green; Franklin Fisher.

[2]For the analysis in the one-sector multifactor framework, see Ryuzo Sato and Tetsunori Koizumi (1973a).

assumed that there are m primary non-produced inputs. Outputs are determined by neoclassical production functions of the form[3]

$$(1) \quad Y_j = F^j(X_{1j}, \ldots, X_{mj}) \qquad j = 1, \ldots, n$$

where Y_j is the output produced and X_{ij} is the input of factor of type i employed in the production of the jth industry. The F^j's are assumed homogeneous of degree one, concave, and twice continuously differentiable. By the homogeneity property, we may write (1)

$$(2) \quad 1 = F^j(a_{1j}, \ldots, a_{mj}) \qquad j = 1, \ldots, n$$

where $a_{ij} = X_{ij}/Y_j$; $i = 1 \ldots m$, $j = 1, \ldots, n$.

Under equilibrium conditions in each sector, we can derive the cost function of that sector as a function of factor prices and output. The cost functions are dual to the production functions, and by the linear homogeneity of the F^j's, can be written as

$$(3) \quad C^j(w_1, \ldots, w_m; Y_j) = Y_j \cdot c^j(w_1, \ldots, w_m)$$

where w_i is the price (or the rental rate) of the ith factor input, C^j is the total cost function, and c^j is the unit-cost function in the jth sector. The function c^j is independent of Y_j and is linearly homogeneous in factor prices.[4] In equilibrium for each industry

$$(4) \quad c^j = p_j \qquad j = 1, \ldots, n$$

where p_j is the price of the jth product, and by a well-known theorem of duality theory,

$$(5) \quad c^j_i = a_{ij}$$

where $c^j_i = \partial c^j/\partial w_i$. Thus, in equilibrium the a_{ij}'s are functions of the factor prices $(w_1, \ldots, w_m)$ only, since there are no returns to scale effects.

To analyze the effect of a change in any of the factor prices on the quantity demanded by each sector of a given input, a convenient form of representing the substitution effects is the (Allen-Uzawa) partial elasticity of substitution, denoted σ^j_{ih}:[5]

$$(6) \quad \sigma^j_{ih} = \frac{1}{S_{hj}} \cdot \frac{\partial \log X_{ij}}{\partial \log w_h} \qquad i, h = 1, \ldots, m, j = 1, \ldots, n$$

where S_{hj} is the relative share of the hth factor in total (optimal) expenditure of the jth sector. The elasticity σ^j_{ih} is calculated for constant output and all other factor prices $\{w_k\}$, $(k \neq h)$. It thus measures the response of the derived demand for the ith factor in a given industry to an input price change, holding output and all other input prices fixed. Under competitive equilibrium conditions[6]

$$(7) \quad \sigma^j_{ih} = \frac{c^j \cdot c^j_{ih}}{c^j_i \cdot c^j_h}$$

where $c^j_{ih} = \partial^2 c^j/\partial w_i \partial w_h$. By the symmetry of the bordered Hessian of the production function F^j we have $c^j_{ih} = c^j_{hi}$ and also $\sigma^j_{ih} = \sigma^j_{hi}$.

In full employment, the aggregate quantities of factor inputs are given by

$$(8) \quad X_i = \sum_{j=1}^{n} X_{ij} \qquad i = 1, \ldots, m$$

Thus, the above quantity and price equations can be expressed in vector notations by

$$(9) \quad X = AY$$

$$(10) \quad p = wA$$

where $X = (X_i)$ is the vector of factors' aggregate quantities, $Y = (Y_j)$ is the vector of outputs, $A = (a_{ij})$ is the matrix of input-output coefficients, $p = (p_j)$ is the vector of output prices and $w = (w_i)$ is the vector of factor prices.

By taking the total differential of (10) we get

$$(11) \quad p' = w'A + wA'$$

where primes denote differentials of the elements of a vector or a matrix. Efficiency conditions imply that

[3]Henceforth, j denotes sectors and i, h, and k denote factor inputs.

[4]See Ronald D. Shephard. The linear homogeneity of the cost function in prices does not depend on the linear homogeneity of F.

[5]See R. G. D. Allen (p. 520).

[6]See Hirofumi Uzawa and Sato and Koizumi (1973b).

$$wA' = 0 \tag{12}$$

To see this, examine the element wa_j' of that vector, a_j being the jth column of A. Under the equilibrium condition (4)

$$a_{ij}' = \sum_{h=1}^{m} c_{ih}^j w_h'; \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \tag{13}$$

Hence

$$wa_j' = \sum_{i=1}^{m} w_i \sum_{h=1}^{m} c_{ih}^j \cdot w_h'$$
$$= \sum_{h=1}^{m} w_h' \sum_{i=1}^{m} c_{ih}^j w_i = 0$$

Since c_i^j is homogeneous of degree zero in prices, and the Euler relation of (5) is given by

$$\sum_{h=1}^{m} c_{ih}^j w_h = \sum_{i=1}^{m} c_{hi}^j w_i = 0$$
$$1 \leq i, \quad h \leq m, \qquad 1 \leq j \leq n$$

The condition $wa_j' = 0$ simply says that the problem "min wa_j; subject to $F^j(a_j) = 1$," has been solved already so that there is no variation in the value of the objective function wa_j for any change in a_j.

In equilibrium in the product markets, the changes in the quantity of output supplied (and demanded) depend on the rate at which output prices are changing and on the price elasticity of the demand for the products. Thus, we assume[7]

$$\hat{Y} = -\eta_D \hat{p} \tag{14}$$

where circumflexes represent relative change or logarithmic partial derivatives of the elements of the vector, and $\eta_D = \text{Diag}(\eta_1, \ldots, \eta_n)$, η_j being the price elasticity of the demand for the jth product, i.e., $\eta_j = -\hat{Y}_j/\hat{p}_j$.

Thus, from (11), (12) and (14),

$$Y' = -w'A \cdot (p_D)^{-1} \eta_D Y_D \tag{15}$$

where $p_D = \text{Diag}(p_1, \ldots, p_n)$ and $Y_D = \text{Diag}(Y_1, \ldots, Y_n)$.

II. Derived Demands and Absolute Shares

The change in the absolute share of a factor as a result of a change in that factor price is directly affected by the derived demand elasticity for that factor. Mathematically stated, if $V_i = w_i X_i$, then

$$\frac{\hat{V}_i}{\hat{w}_i} = \frac{\hat{w}_i + \hat{X}_i}{\hat{w}_i} = 1 + \frac{\hat{X}_i}{\hat{w}_i} = 1 - \lambda_i$$

where λ_i is the derived demand elasticity of the ith factor input. Since the ith factor input is used in n production processes it would be interesting to examine how λ_i is related to the other demand and supply elasticities.

By taking the total differential of (9) we get $X' = A'Y + AY'$ which, by (15) implies

$$X' = A'Y - A(p_D)^{-1} \eta_D Y_D A^T w' \tag{16}$$

where superscript T denotes the transpose of a matrix (vector).

When the only change is in the price of the hth factor, the resulting change in the aggregate derived demand for the ith factor is obtained from (16), and via (14) and (7) can be expressed in elasticity terms by

$$\frac{\partial X_i}{\partial w_h} = \left[\sum_{j=1}^{n} a_{ij} a_{hj} \sigma_{ih}^j \cdot \frac{Y_j}{p_j}\right] - \left[\sum_{j=1}^{n} a_{ij} a_{hj} \eta_j \cdot \frac{Y_j}{p_j}\right]$$

which can also be written as

$$\frac{\partial X_i}{\partial w_h} = \sum_{j=1}^{n} \frac{X_{ij} \cdot X_{hj}}{P_j Y_j} (\sigma_{ih}^j - \eta_j) \tag{17}$$

[7] Notice that in assuming zero cross-price elasticities and disregarding the income effect on demand of a change in input prices, the implicit assumption is that the individuals in the economy are momentarily *not* in equilibrium. Thus, the cross-price effects as well as the income effect are not shown in the markets over the short run. We, therefore, examine the initial response of the producing sectors to an exogeneous shift in factor prices along their short-run expansion path. This is in line with the traditional Hicksian analysis of the derived demand for factor inputs where interest is centered on the *initial* effect on the quantity demanded of a factor resulting from a change in its own price or in other prices. According to this assumption the producers' response to factor-price shifts is initially based on the change in their production costs and on the demand conditions for their own products. In revising the production plans and thus the derived demand for factor inputs, other effects—which would be reflected in the markets over a longer period of time—are not initially taken into account. It should be noted, however, that the paper's main conclusions remain intact even when the cross-price effects are taken into account.

As a special case of (17) one gets the relation

$$\frac{\partial \log X_i}{\partial \log w_h} = S_h(\sigma_{ih} - \eta)$$

derived in the one-product many-inputs model by Allen, p. 508.

When *all* factor prices are changing, the relative change in the derived demand for the ith factor input, is given by

$$(18)\quad \hat{X}_i = \sum_{j=1}^{n}\sum_{h=1}^{m} \frac{R_j}{S_i}\cdot S_{ij}\cdot S_{hj}\,(\sigma^j_{ih} - \eta_j)\cdot \hat{w}_h$$

where $R_j = p_j\,Y_j/pY$ is the share of the jth industry in total national output, $S_i = w_iX_i/pY$ is the share of the ith factor in total national income, and $S_{ij} = w_iX_{ij}/p_jY_j$ is its share in the income of the jth sector. The one-sector multifactor analog of (18) is given (omitting the subscript j) by

$$(18')\qquad \hat{X}_i = \sum_{h=1}^{m} S_h(\sigma_{ih} - \eta)\cdot \hat{w}_h$$

The relative change in the absolute share of a factor as an effect of a change in any one of the factor prices is directly obtained from (17). Denote the absolute share of the ith factor input by V_i, then $V_i = w_iX_i$ and $\hat{V}_i = \hat{w}_i + \hat{X}_i$. Thus, when $h \neq i$,

$$(19)\quad \frac{\partial \log V_i}{\partial \log w_h} = \sum_{j=1}^{n}\frac{R_j}{S_i}\,S_{ij}S_{hj}(\sigma^j_{ih} - \eta_j)$$

and when $h = i$, i.e. the change is in its own price,

$$(19')\quad \frac{\partial \log V_i}{\partial \log w_i} = \sum_{j=1}^{n}\frac{R_j}{S_i}\,S_{ij}\cdot\Big[\sum_{h\neq i}^{n} S_{hj}(1 - \sigma^j_{ih}) + S_{ij} - 1 - \eta_j\Big]$$

Equation (19′) shows how the derived demand elasticity λ_i is related to the other elasticities in a special case where supply of all other inputs is perfectly inelastic. The one-sector multifactor analog of (19′) is given by

$$(19'')\quad \frac{\partial \log V_i}{\partial \log w_i} = \sum_{h\neq i}^{n} S_h(1 - \sigma_{ih}) - S_i\eta$$

Obviously, if the internal structure of the production function on the one hand and the demand conditions for the product of each sector on the other are such that an increase in the price of one factor results in an increase (or a decrease) in the absolute share of that factor in each of the sectors in the economy, the same results will be recorded for the aggregate absolute share of that factor. This, however, is no longer the case with regard to the *relative* share of a factor, which we now turn to examine.

III. Relative Shares

To derive the change in the relative share of a factor as an effect of a change in any one of the factor prices take the logarithmic partial differential of $S_i = V_i/pY$ to get

$$(20)\qquad \hat{S}_i = \hat{V}_i - \frac{p'Y + pY'}{pY}$$

From the equilibrium condition $pY = wX$, we have

$$(21)\qquad p'Y + pY' = w'X + wX'$$

From (16) we get

$$(22)\quad wX' = wA'Y - wA(p_D)^{-1}\eta_D(Y_D)A^Tw'$$

which by (10) and (12) yields

$$(23)\qquad wX' = -\,w'A\eta_D Y$$

By referring to the Walras identity $wX = pY$ and the condition $wA' = 0$ (equation (12)) we get

$$\begin{aligned} w'X + wX' &= w'AY + wA'Y + wAY' \\ &= w'X + (wA')Y + pY' \\ &= w'X + pY' \end{aligned}$$

Hence

$$(24)\qquad wX' = pY'$$

Heuristically equation (24) says that the economy-wide profit-maximization problem "max $(pY - wX)$" has been solved by optimally adjusting the Y and X vectors.

Thus the real change in the physical inputs evaluated or weighted by the equilibrium factor prices (rental rates) is equal to the real change in total output, defined as the changes in the physical quantities pro-

duced evaluated at the equilibrium prices of the products. By referring to equations (23) and (24) we can observe that as an effect of an increase in factor prices there will *always* be a real decline in the level of income and output. This, however, need not be the case with regard to the *value* of income and output because of the price effect, since

$$(25) \quad w'X + wX' = w'A(I - \eta_D)Y$$

When the only change is in the price of the hth factor input, (25) can be written as

$$(26) \quad \frac{w'X + wX'}{wX} = \sum_{j=1}^{n} R_j S_{hj}(1 - \eta_j) \cdot \hat{w}_h$$

Thus, from (19), (20), and (26), the change in the relative share of the ith factor input as an effect of a change in the price of the hth factor input, $h \neq i$, is given by

$$(27) \quad S_i' = \sum_{j=1}^{n} R_j S_{hj}[S_{ij}(\sigma_{ih}^j - \eta_j) - S_i(1 - \eta_j)] \cdot \hat{w}_h$$

and the effect of a change in its own price, i.e., $h = i$, is given by

$$(27') \quad S_i' = \sum_{j=1}^{n} R_j S_{ij}[\sum_{h \neq i} S_{hj}(1 - \sigma_{ih}^j) - (S_i - S_{ij})(1 - \eta_j)] \cdot \hat{w}_i$$

The one-sector multifactor analog of (27) is given in relative change terms by

$$(27'') \quad \hat{S}_i = S_h(\sigma_{ih} - 1) \cdot \hat{w}_h$$

From (27″) one can derive the relations, presented by Sato and Koizumi (1973a, equation 6a), between the sign of $\partial S_i/\partial w_h$ and the distribution of the corresponding elasticity of substitution on either side of unity. Notice that in the one-sector framework, the price elasticity of demand does not affect the change in the relative share, whereas it does have an effect in the multisector framework. In the one-sector case this is due to the absence of a returns to scale effect, whereby the relative share of a factor is determined by variables which are independent of the level of output, i.e., $S_i = a_i w_i / c$. By taking the logarithmic total differential of the latter expression, together with equilibrium conditions (5) and (7), we obtain the rate of change in the relative share of the ith factor input as an effect of the change in *all* factor prices.

$$(27''') \quad \hat{S}_i = \sum_{h=1}^{n} S_h(\hat{w}_i - \hat{w}_h)(1 - \sigma_{ih})$$

The latter structural equation provides, as special cases, the other relations presented by Sato and Koizumi for exogenous price shifts. From (27″) we conclude that the relative share of the ith factor input in total (optimal) expenditure of a given sector will increase, remain constant, or decrease as an effect of an increase in the price of the hth factor input according as the corresponding elasticity of substitution between these two factors is greater than, equal to, or smaller than unity, i.e., $sgn \partial S_i/\partial w_h \gtreqless 0$ according as $\sigma_{ih} \gtreqless 1$. Yet in (27) we can see that even when this condition is satisfied for each of the n sectors in the economy it is neither necessary nor sufficient to ensure that the aggregate share of that factor in total national income will also change in the same direction. The reason for that is the simultaneous changes in output prices and quantities and thus the changes in the value of the national income, that do have an effect on the aggregate share of the factor.

(By 27″), we can write (27) as

$$(28) \quad S_i' = \sum_{j=1}^{n} R_j S_{ij}' + \sum_{j=1}^{n} R_j S_{hj}(S_i - S_{ij}) \cdot (\eta_j - 1) \cdot \hat{w}_h$$

The first expression on the right-hand side of (28) is a weighted average of the changes in sectoral shares. The second expression manifests the effect of expansion (or contraction) in output. Solow has counted only the first expression in discussing the aggregation of changes in the relative shares at the industry levels up to the national level. In (28) we observe, however, that the expansion effect might be quite significant, unless all demand elasticities are identically unity.

REFERENCES

R. G. D. Allen, *Mathematical Analysis for*

Economics, London 1938.

F. M. Fisher, "The Existence of Aggregate Production Functions," *Econometrica*, Oct. 1969, *37*, 553-77.

H. A. J. Green, "Embodied Progress, Investment and Growth," *Amer. Econ. Rev.*, Mar. 1966, *56*, 138-51.

John R. Hicks, *The Theory of Wages*, 2d ed., London 1963.

I. B. Kravis, "Relative Income Shares in Fact and Theory," *Amer. Econ. Rev.*, Dec. 1959, *49*, 917-49.

R. Sato, "The Estimation of Biased Technical Progress and Production Function," *Int. Econ. Rev.*, June 1970, *11*, 179-208.

——— **and T. Koizumi**, "Substitutability, Complementarity and the Theory of Derived Demand," *Rev. Econ. Stud.*, Jan. 1970, *37*, 107-18.

——— **and** ———, (1973a) "The Production Function and the Theory of Distributive Shares" *Amer. Econ. Rev.*, June 1973, *63*, 484-89.

——— **and** ———, (1973b) "On the Elasticities of Substitution and Complementarity," *Oxford Econ. Pap.*, Mar. 1973, *25*, 44-56.

Ronald W. Shephard, *Theory of Cost and Production Function*, Princeton 1970.

R. M. Solow, "A Skeptical Note on the Constancy of Relative Shares," *Amer. Econ. Rev.*, Sept. 1958, *48*, 618-31.

A. Takayama, "On Biased Technological Progress," *Amer. Econ. Rev.*, Sept. 1974, *64*, 631-39.

M. Uzawa, "Neutral Inventions and the Stability of Growth Equilibrium," *Rev. Econ. Stud.*, Feb. 1961, *28*, 117-24.

Determining the Monetary Instrument: A Diagrammatic Exposition

By STEPHEN F. LEROY AND DAVID E. LINDSEY*

The problem of determining short-run monetary policy is often posed as that of choosing which of several variables to take as the monetary instrument, which is understood to mean choosing which variable to maintain at a preassigned level under random shifts in the structural equations. In the simplest case, this problem has been unambiguously solved. Suppose that we have a static linear *IS-LM* structure with independent normally distributed errors and known coefficients:

$$y = a_0 + a_1 i + u \qquad a_1 < 0$$
$$y = b_0 + b_1 m + b_2 i + v \qquad b_1 > 0, b_2 > 0$$

where y, i, and m are income, the interest rate, and the money stock, respectively, and the errors u and v have zero means and variances σ_u^2 and σ_v^2.[1] Assume that the loss equals the mean squared deviation of income around some target level y^* and that the money stock and the interest rate (but not income) are observable, implying that the monetary authority can adopt either a pure interest rate or a pure money stock policy. Its choice between these will depend on which generates a lower loss when set at its optimal level.

It can be shown that under either instrument, the loss equals the variance of the error in the reduced-form equation for income when that instrument is taken as exogenous.[2] Since the reduced form for income is

$$y = a_0 + a_1 i + u$$

when the interest rate is the instrument (i.e., it coincides with the *IS* equation) and

$$y = \frac{a_0 b_2 - a_1 b_0 - a_1 b_1 m + b_2 u - a_1 v}{b_2 - a_1}$$

under a money stock policy, the interest rate is taken as the instrument if

$$(1) \qquad \sigma_u^2 < \frac{b_2^2 \sigma_u^2 + a_1^2 \sigma_v^2}{(b_2 - a_1)^2}$$

and the money stock is taken as the instrument if the reverse inequality is satisfied.

The analysis of instrument choice just presented, due originally to Poole (1970), has been accorded widespread attention. It has been applied in empirical studies of the choice both of an intermediate target (see Robert Holbrook and Harold Shapiro) and

*University of California-Santa Barbara and Federal Reserve Board, respectively. The views expressed here are solely our own and do not necessarily represent the views of the Board of Governors of the Federal Reserve System.

[1]The assumption that the errors u and v are independent is a normalization and involves no loss of generality. To see this, observe that if $\sigma_{uv} \neq 0$, a new *LM* equation may be derived by multiplying the *IS* equation by σ_{uv}/σ_u^2, subtracting the result from the *LM* equation and solving for income. The new *LM* equation is indistinguishable from the original equation, except that its error is independent of u. In econometric terms, the normalization $\sigma_{uv} = 0$ may be interpreted as a restriction serving to identify the structural parameters. For a discussion of identification from covariance restrictions, and identification generally, see Franklin M. Fisher, especially ch. 4. The *LM* function is solved for y rather than m because it will be convenient below to have the errors u and v expressed in common units.

[2]See William Poole (1970) where the solution was first presented, and LeRoy and Roger Waud (1977) for further discussion. Poole also discusses a "combination policy" under which the two candidate variables are assumed to be linearly related. In LeRoy and Waud it is argued that the monetary authority has a choice among instruments only if more than one variable is currently perfectly observable, and in that case the monetary authority would always choose the optimal combination policy in preference to a pure policy (unless, of course, the combination policy coincides with one of the pure policies). The primary focus of this article on pure policies is justified by the fact that these have dominated professional discussion of the instrument problem to the virtual exclusion of the combination policy, even though the latter has a sounder analytical grounding.

of an operating target for controlling money (see, for example, James Pierce and Thomas Thomson). Poole's analysis has recently been invoked by both sides in the debate regarding the policy implications of the alleged shift in the money demand function (see, for example, Jared Enzler, Lewis Johnson, and John Paulus; Stephen Goldfeld; Michael Hamburger). Poole (1971) also presented a diagrammatic exposition of instrument choice which has become well known. It is not commonly realized that his diagrammatic representation of instrument choice is based on assumptions different from, and less appealing than, those underlying the algebraic treatment, and that the two methods sometimes generate opposite recommendations for the monetary instrument (see our earlier version of this paper). Our purpose in this article is to present a diagrammatic analysis of instrument choice that is based on, and is therefore consistent with, Poole's algebraic treatment. We also derive implications of the analysis which have not been recognized before. The development in the text is at an intuitive level; a technical justification for the results is found in the Appendix.

The proposed diagrammatic development is based on the fact that in the model under discussion both the interest rate and the money stock are observable. (This must be true since otherwise the Federal Reserve would not be able to choose either of the two variables as the monetary instrument.)[3] The fact that both variables are observable means that information is available upon which to base conditional estimates of the random shifts in the IS and LM equations and, therefore, upon which to appraise the likely effect of either pure policy on income. Specifically, in Figure 1 let IS_1 and LM_1 be the normal IS and LM functions, intersecting at $y = y^*$ and $i = i^*$. They are drawn for $u = v = 0$, and the LM curve is drawn for the certainty equivalence value of money m^*, implied by y^* and i^*. Suppose, however, that in the current period the Federal

[3]See LeRoy, and LeRoy and Waud (1976, 1977) for discussion of the notion of observability.

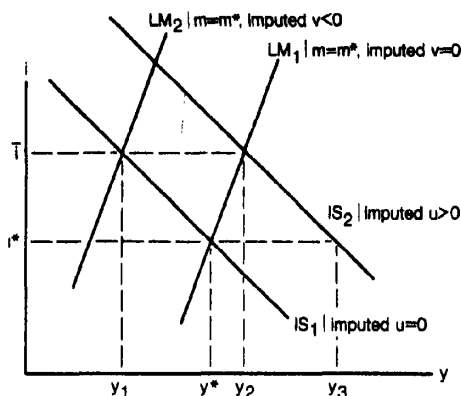



FIGURE 1

Reserve discovers that if it maintains the money stock at m^*, the interest rate will rise to $\bar{i}$, higher than the normal i^*, in response to the realizations of u and v. The Federal Reserve needs to determine the likely effect on income if it, in fact, allows this rise in the interest rate to occur by maintaining the money stock at m^* vs. that which would occur if it were to stabilize the interest rate at the normal level i^*, by instead allowing the money stock to increase. To make this determination, it is necessary to estimate whether the IS curve has shifted to the right, due to a positive realization of u, or the LM curve has shifted to the left, due to a negative realization of v, or both. Because income is not observable, the separate realizations of u and v cannot be determined directly and must be imputed. These estimates will necessarily depend on the relation between the error variances of the respective functions. If the IS equation is completely stable ($\sigma_u^2 = 0$), the entire variation in the interest rate is imputed to the LM curve, which must have shifted to LM_2 if $m = m^*$ and $i = \bar{i}$ are to be consistent. Accordingly, income will equal y_1 under a money stock policy and y^* under an interest rate policy, since the normal interest rate i^* can be maintained only if enough money is supplied to shift the LM function back to LM_1. Obviously, an interest rate policy is preferred, since under the assumed stability of the IS equation an

interest rate policy guarantees that income will always equal its optimal level. Now, consider the reverse situation in which the *LM* equation is completely stable ($\sigma_v^2 = 0$). In this case, the entire shift in the interest rate is imputed to the *IS* curve, implying that income will equal y_2 under a money stock policy and y_3 under an interest rate policy. In either case, income exceeds the optimum, but the excess is greater under an interest rate policy, implying that a money stock policy is preferred.

Now suppose that both the *IS* and *LM* equations are subject to random shifts. The analysis of the simpler case in which either the *IS* or the *LM* equation is stable suggests that if both equations are unstable, the change in the interest rate which would occur under a money stock policy will be imputed to shifts in both curves. On intuitive grounds, one would expect the amount of the shift imputed to each equation to be proportional to the variance of its error relative to that of the other, and this turns out to be exactly the case.

We have just shown that if the interest rate equaled $\bar{i}$ and the *IS* (*LM*) curve were stable, expected income would be expressed by the horizontal coordinate of the intersection of the IS_1 (LM_1) equation with the line $i = \bar{i}$, denoted by *A* (*B*) in Figure 2. If both equations are subject to random errors, we may determine the imputed shifts in the *IS* and *LM* curves by defining a point *C* on the line segment *AB* such that the length of the segment *AC* is to the length of *CB* as the variance of *u* is to the variance of *v*. In Figure 2, it is assumed that σ_v^2 is about twice as great as σ_u^2; consequently *C* is chosen about one-third of the distance from *A* to *B*. Now, construct IS_2 and LM_2 parallel to IS_1 and LM_1 but passing through *C*. These lines represent the imputed shifts in the *IS* and *LM* curves. Accordingly, the horizontal coordinate of point *C* represents expected income under a money stock policy (y_m), while the horizontal coordinate of the intersection of IS_2 with $i = i^*$ represents expected income under an interest rate policy (y_i). In the current example, each differs from y^* by about an equal amount, implying that the Federal Reserve will be approximately indifferent between a money stock policy and an interest rate policy.

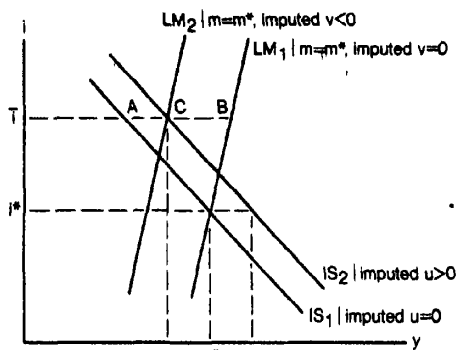



FIGURE 2

The diagrammatic procedure just outlined can be used to develop an intuitive motivation of the optimal combination policy as well as to analyze pure policies. Given that the Federal Reserve knows that maintaining either the money stock or the interest rate will result in conditional bias, we can ask why should the Federal Reserve not move the money stock (or, equivalently, the interest rate) so as to eliminate the bias. In Figure 3, suppose that under an unchanged money stock an interest rate $i = \bar{i}$ would result, and suppose also that IS_2 and LM_2 are the imputed *IS* and *LM* functions, as before. Rather than maintaining the

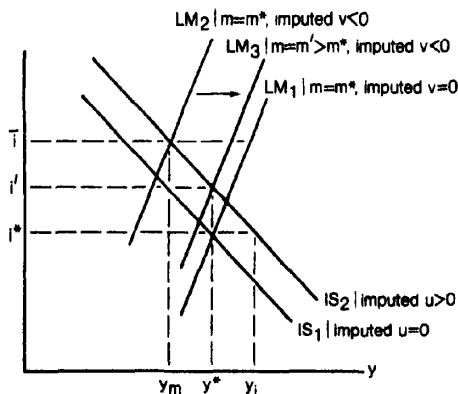



FIGURE 3

money stock unchanged, which results in an interest rate that is too high and expected income that is too low, or maintaining the normal interest rate, and thus causing money and expected income to be too high, the Federal Reserve can adopt an intermediate position. If the Federal Reserve increases the money stock to m' and lowers the interest rate to i' by shifting the imputed LM curve to LM_3 (where m' is such that LM_3 intersects IS_2 at $y = y^*$), the conditional bias in monetary policy will be eliminated. It can be shown that this procedure is equivalent to a combination policy in which the interest rate and money stock are maintained in the particular linear combination policy which minimizes the expected loss.

Experimenting with the diagrams reveals several instructive special cases. If the error variances of the IS and LM equations are equal, a pure money stock policy is always preferred to a pure interest rate policy for any admissible values of the equations' slopes. Moreover, equality of the equations' error variances combined with algebraic equality of the equations' slopes is sufficient for the optimum combination policy to coincide with a pure money stock policy.

Finally, it is worthwhile noting that our diagrammatic construction cannot be used to evaluate the effect on the loss of changes in error variances; failure to realize this will result in invalid conclusions. For example, it might seem plausible that under a pure money stock policy which has some bias, the expected loss can be decreased by increasing the variance of one of the structural errors so that the horizontal coordinate of the intersection of IS_2 and LM_2 would occur closer to y^*. Obviously, the actual outcome of such a comparative statics experiment would be that the expected loss would rise. The reason the diagrammatic argument is improper is that the combination policy loss, which is not represented in the diagrams but which is a component of the total expected loss, is affected by a change in the error variances. Additionally, the presumption that a given $\bar{i}$ can be maintained under a change in the error variances renders the diagrams meaningless since the variance of $i - i^*$ is a function of these error variances.

Appendix

To justify the diagrammatic method presented in the text for choosing between pure policies, we first derive expressions for the conditionally expected deviations of income from the optimum under either pure policy. Then it is shown that these expressions can be viewed as forming the basis for the imputed shifts in the IS and LM functions constructed in Figure 2. Finally, the proof of the validity of the diagrammatic procedure consists in showing that an inequality between the expressions for the respective deviations is equivalent to inequality (1) of the text. In the remainder of the Appendix, a somewhat more abstract interpretation of the diagrammatic construction is developed.

We begin by determining the expectation of u and v conditional on $i = \bar{i}$ under a money stock policy. First, write the reduced-form equations for y and i:

$$\text{(A1)} \quad y = \frac{a_0 b_2 - a_1 b_0 - a_1 b_1 m + b_2 u - a_1 v}{b_2 - a_1}$$

$$\text{(A2)} \quad i = \frac{a_0 - b_0 - b_1 m + u - v}{b_2 - a_1}$$

These expressions may be simplified if y and i are expressed as deviations from y^* and i^*. To do this, set $u = v = 0$ and $y = y^*$, calculate the implied value for m from (A1), and define i^* as the value of i determined by entering this value for m in (A2). Then (A1) and (A2) may be written as

$$\text{(A3)} \quad y = y^* + \frac{b_2 u - a_1 v}{b_2 - a_1}$$

$$i = i^* + \frac{u - v}{b_2 - a_1}$$

We may now determine the conditional expectations of the structural errors in the IS and LM functions. These may be shown to equal[4]

[4]Here and elsewhere in the Appendix we make use of the fact that if

$$(A4)\quad u^e \equiv E(u \mid i = i^* + \frac{u - v}{b_2 - a_1} = \bar{i}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}(b_2 - a_1)(\bar{i} - i^*)$$

$$v^e \equiv E(v \mid i = i^* + \frac{u - v}{b_2 - a_1} = \bar{i}) = -\frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2}(b_2 - a_1)(\bar{i} - i^*)$$

The fact that the estimates of realized u and v are proportional to their respective variances justifies the construction of IS_2 and LM_2 in Figure 2 reported in the text in terms of conditional expectations, since AC and CB are just u^e and $-v^e$, respectively. By substituting (A4) in the reduced forms for y under an interest rate and money stock policy, we have

$$y_i = y^* + \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}(b_2 - a_1)(\bar{i} - i^*)$$

$$y_m = y^* + \frac{b_2\sigma_u^2 + a_1\sigma_v^2}{\sigma_u^2 + \sigma_v^2}(\bar{i} - i^*)$$

Thus, our diagrammatic procedure implies that an interest rate policy is preferred if

$$(A5)\quad \left| \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}(b_2 - a_1)(\bar{i} - i^*) \right| < \left| \frac{b_2\sigma_u^2 + a_1\sigma_v^2}{\sigma_u^2 + \sigma_v^2}(\bar{i} - i^*) \right|$$

and a money stock policy is preferred under the reverse inequality. It remains to show that this criterion coincides with that based on the variance of the reduced-form errors. If both sides of (A5) are squared and common terms are cancelled, we have

$$\sigma_u^4 < \frac{(b_2\sigma_u^2 + a_1\sigma_v^2)^2}{(b_2 - a_1)^2}$$

Now, if we add $\sigma_u^2\sigma_v^2$ to both sides, simplify the right-hand side and divide both sides by $\sigma_u^2 + \sigma_v^2$, we obtain

$$\sigma_u^2 < \frac{b_2^2\sigma_u^2 + a_1^2\sigma_v^2}{(b_2 - a_1)^2}$$

agreeing with (1) as required. All these operations are reversible, so the equivalence is proved.

The logic of this proof may be clarified if we examine the diagrammatic construction from a more general and abstract viewpoint. The laws of probability allow us to write the expected loss $\underset{u,v}{E}(y - y^*)^2$ in the form

$$(A6)\quad \text{expected loss} = \underset{i}{E}[\underset{u,v|i}{E}(y - y^*)^2]$$

that is, we can calculate the loss by taking the expectation of $(y - y^*)^2$ over u and v conditional on u and v being such as to generate a particular interest rate under a money stock policy, and then taking expectations over the distribution of the interest rate. Further, we can break up the expected loss conditional on the interest rate into two parts by use of an identity. Accordingly, we have

$$(A7)\quad \text{expected loss} = \underset{i}{E}\{\underset{u,v|i}{E}[y - \underset{u,v|i}{E}(y)]^2 + [\underset{u,v|i}{E}(y) - y^*]^2\}$$

Here we have expressed the mean squared deviation of y from y^* in (A6) as the sum of the conditional variance and the squared conditional bias, an identity familiar to econometricians. Now, because of the linearity of the model, the variance of income conditional on the interest rate is independent of the interest rate (see fn. 4), and hence can be taken out of the braces. Since the variance equals $\sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2)$, we have

$$\text{expected loss} = \frac{\sigma_u^2\sigma_v^2}{\sigma_u^2 + \sigma_v^2} + \underset{i}{E}\{[\underset{u,v|i}{E}(y) - y^*]^2\}$$

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ has mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

then Ax has mean $A\mu$ and covariance matrix $A\Sigma A'$, and the distribution of x_1 conditional on x_2 has mean

$$\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$$

and variance $\sigma_{11} - (\sigma_{12}^2/\sigma_{22})$.

Also by linearity, the conditional bias will be proportional to $i - i^*$, so we have finally

$$\text{expected loss} = \frac{\sigma_u^2 \sigma_v^2}{\sigma_u^2 + \sigma_v^2} + E_i\{[\lambda(i - i^*)^2]\} = \frac{\sigma_u^2 \sigma_v^2}{\sigma_u^2 + \sigma_v^2} + \lambda^2 \, var(i)$$

The parameter λ thus gives the conditional bias as a function of the interest rate. It is evident that the value of λ will be different for an interest rate policy than under a money stock policy and also that an interest rate policy is preferred to a money stock policy if and only if the absolute value of λ is lower in one than in the other. Now, it is seen that the diagrammatic procedure for choosing between pure policies consists of developing a geometric construction of λ under either pure policy, as was just proved. Incidentally, the term $\sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2)$, the variance of income conditional on the interest rate, may be recognized as simply the expected loss under an optimal combination policy (see LeRoy; LeRoy and Waud, 1977; Poole, 1970).

REFERENCES

J. Enzler, L. Johnson, and J. Paulus, "Some Problems of Money Demand," *Brookings Papers*, Washington 1976, *1*, 261–80.

Franklin M. Fisher, *The Identification Problem in Econometrics*, New York 1966.

S. M. Goldfeld, "The Case of the Missing Money," *Brookings Papers*, Washington 1976, *3*, 683–730.

M. J. Hamburger, "Behavior of the Money Stock: Is There A Puzzle?," *J. Monet. Econ.*, July 1977, *3*, 265–88.

R. Holbrook and H. Shapiro, "The Choice of Optimal Intermediate Economic Targets," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 40–46.

S. F. LeRoy, "Efficient Use of Current Information in Short-Run Monetary Control," spec. stud. paper 66, Fed. Res. Board, Washington 1975.

______ **and D. E. Lindsey,** "Determining the Monetary Instrument: A Diagrammatic Exposition," spec. stud. paper 103, Fed. Res. Board, Washington 1977.

______ **and R. N. Waud,** "Observability, Measurement Error and the Optimal Use of Information for Monetary Policy," spec. stud. paper 72, Fed. Res. Board, Washington 1976.

______ **and** ______, "Applications of the Kalman Filter in Short-Run Monetary Control," *Int. Econ. Rev.*, Feb. 1977, *18*, 195–207.

W. Poole, "Optimal Choice of Monetary Policy in a Simple Stochastic Macro Model," *Quart. J. Econ.*, May 1970, *84*, 197–216.

______, "Rules-of-Thumb for Guiding Monetary Policy," in *Open Market Policies and Operating Procedures—Staff Studies*, Fed. Res. Board, Washington 1971, 137–89.

J. L. Pierce and T. D. Thomson, "Some Issues in Controlling the Stock of Money," in *Controlling Monetary Aggregates II: The Implementation*, Fed. Res. Bank Boston Conference Series, No. 9, Sept. 1972, 115–36.

Do Managers Use their Information Efficiently?

By STEVEN SHAVELL*

It is often true that a manager's opinions about events relevant to production are valued but are not fully known by others. This note suggests that in such circumstances there may be a problem with production. Consider a competitive equilibrium in a standard Arrow-Debreu model of an economy. In such an equilibrium production decisions are guided by prices and, in particular, by contingent commodity prices (which in fact may be implicit in stock market prices). Moreover, in such an equilibrium the managers of production processes play a strictly passive role since complete instructions for production are implicit in the criterion of profit maximization.[1] However, if the probabilistic beliefs of the managers are valued but are not fully known by the other agents in the economy, then it seems that these agents might well prefer to have the managers play an active role in making production decisions. In other words, it seems that profit maximization with respect to contingent commodity prices may encourage managers to act contrary to what would be the best wish of others, and consequently that the absence of markets in certain contingent commodities might not be undesirable.[2]

Our discussion of this issue will make reference to a simple example. An economy with many identical individuals and few identical managers uses seed to produce wheat which may be grown in two regions, A and B. Managers decide where to plant the seed. The wheat harvest is uncertain—it is either positive or zero—depending on which of the two possible states of nature, α and β, occurs. This is described in Table 1, where s_i is the amount of seed planted in region i and f is the usual type of production function ($f' > 0, f'' < 0$). Let us suppose for simplicity that consumers alone determine prices in competitive equilibrium, that is, the few managers have only a negligible impact on the prices.

Assume initially that consumers have fixed beliefs, independent of those which the managers might have. Specifically, assume that consumers believe the state α will occur with probability a. Then, since a competitive equilibrium in which there are markets for contingent wheat is Pareto efficient, it must in this case maximize expected utility of consumers. Consequently, if each consumer's endowment consists of one unit of seed and his von Neumann-Morgenstern concave utility function $U(\cdot)$ depends only on consumption of wheat, the problem solved by the market is to maximize expected utility:

$$(1)\quad \max_{s_A \in [0,1]} aU(f(s_A)) + (1-a)U(f(1-s_A))$$

As one would guess, the first-order condition (which shall be assumed to hold)

$$(2)\quad \frac{a}{1-a} = \frac{f'(1-s_A)U'(f(1-s_A))}{f'(s_A)U'(f(s_A))}$$

implies that the higher the probability a of the state α, the more seed s_A is planted in region A. The beliefs of profit-maximizing managers do not affect the amount of seed planted in the two regions;[3] given the prices

*Harvard University. I wish to thank P. Diamond, G. Feiger, A. M. Polinsky, and L. Weiss for comments and the National Science Foundation (grant no. SOC-76-20862) for research support.

[1]See Kenneth Arrow.

[2]This point may be compared with those in Jack Hirshleifer's important article, which emphasizes matters related to investment in and dissemination of information. Here there is no investment in information nor is there any transfer of information among agents in the economy; the stress is instead on logically distinct questions concerning the use of information as it exists in the economy, information "in place."

[3]As is well known, in the situation described managers would be instructed by consumer-stockholders to maximize profits (or, equivalently, stock market value).

TABLE 1—WHEAT HARVEST

Region	State of Nature α	β
A	$f(s_A)$	0
B	0	$f(s_B)$

of seed and of contingent wheat, managers' beliefs influence only their purchases of contingent wheat (for their own consumption).

Now assume that the consumers' beliefs would be influenced by the managers', if only they knew them. Assume furthermore that consumers learn nothing about managers' beliefs.[4] Let $a_1 > a_2$ be the two probabilities of the state α which consumers believe could possibly characterize the opinions of the managers. Let μ be the consumers' probability that the managers' probability is a_1 and let a^1 and a^2 be the consumers' probabilities of α conditional on the managers' probabilities of a_1 and a_2, respectively. Then the consumers' probability a of α satisfies

$$a = \mu a^1 + (1 - \mu)a^2 \tag{3}$$

which is to say that the consumers' beliefs are a weighted average of what they would be, given the two possibilities for those of the managers. If managers are regarded as experts by consumers, then a^1 would be close to a_1 and a^2 to a_2. If the consumers believed in a theory of "contrary opinion," it might be that $a^1 < a^2$. As long as $a^1 \neq a^2$ and μ is not zero or one, the consumers would wish to know the managers' beliefs. In general, it seems natural to say that one individual's probabilistic beliefs are *valuable* to another if the latter thinks that there is a positive probability that his probabilistic beliefs would change on revelation of those of the first.

Assuming that the manager's beliefs are valuable to consumers, it is clear that the consumers might be made better off if production were affected by the managers' beliefs. To make consumers as well off as possible, a benevolent dictator would solve the following problem.

$$\max_{s_{Ai}\in[0,1]} \mu[a^1 U(f(s_{A1})) + (1 - a^1)U(f(1 - s_{A1}))] + (1 - \mu)[a^2 U(f(s_{A2})) + (1 - a^2)U(f(1 - s_{A2}))] \tag{4}$$

where s_{A1} is the amount of seed planted in region A if the manager's probability of α is a_1 and s_{A2} is defined similarly. This problem is obviously different from (1), what the market solves. Suppose, for example, that $a^1 > a > a^2$—the consumers would change their beliefs in the direction of the managers'. Then the benevolent dictator would order the manager to plant more in region A if his probability is a_1 than if it is a_2.[5]

This notional inefficiency might be viewed in a purely formal way as arising from a lack of markets—in wheat contingent on pairs, one element of which is the usual state of nature and the other, the probability distribution of the manager. Markets in wheat contingent on such pairs are hard to imagine, especially because of consumers' difficulty in verifying the "occurrence" of the managers' probability distribution.

In actual fact, the absence of markets in many contingent commodities—contingencies now being taken in the usual sense of states of nature—may mitigate the type of inefficiency of concern here. This is because the absence of markets for commodities in certain contingencies means that managers are not given (implicitly) complete instructions for operating the production process by the market. Instead they are forced to make decisions which are influenced by their own beliefs, and are therefore of potential benefit to consumers.

Let us conclude by illustrating how the

[4]As will be clear, only the presence of valuable opinion—of less than full communication of beliefs—needs to be assumed. It therefore does not seem crucial that we have ignored how consumers might in fact learn something about managers' beliefs.

[5]From (4), it is clear that the optimal s_{Ai} maximizes $a^i U(f(s_{Ai})) + (1 - a^i)U(f(1 - s_{Ai}))$. But from (2), we know that the solution to this is increasing in a^i.

absence of markets in contingent commodities may improve matters. Suppose in our example that there is only a market in wheat, not a market in wheat contingent on the two possible states of nature. Suppose further that the managers have the same utility function as consumers and act so as to maximize their expected utility, where the expectation is with respect to their probability distribution.[6] Then the consumers' expected utility is given by

$$(5)\quad \mu[a^1 U(f(s_{A1})) + (1 - a^1)U(f(1 - s_{A1}))] + (1 - \mu)[a^2 U(f(s_{A2})) + (1 - a^2)U(f(1 - s_{A2}))]$$

where the s_{Ai} are chosen by the managers so as to

$$(6)\quad \max_{s_{Ai}\epsilon[0,1]} a_i U(f(s_{Ai})) + (1 - a_i)U(f(1 - s_{Ai}))$$

[6]If managers are given an appropriate share of total wheat output, this is what they would wish to maximize.

It is clear that consumers might be better off with expected utility determined by (5) and (6) rather than by (1). For example, if $a^1 = a_1$ and $a^2 = a_2$—the consumers would agree with the managers' beliefs if they knew them—then consumers would enjoy exactly the expected utility achieved under the benevolent dictator's first best solution.[7]

[7]For under this assumption, expected utility determined by (4) and by (5) and (6) are identical.

REFERENCES

K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk-Bearing," in his *Essays in the Theory of Risk-Bearing*, Chicago 1971, 131-33.

J. Hirshleifer, "The Private and Social Value of Information and the Reward to Inventive Activity," *Amer. Econ. Rev.*, June 1971, *61*, 561-74.

Cartel Problems: Comment

By DAVID E. MILLS AND KENNETH G. ELZINGA*

In a recent paper in this *Review*, Dale Osborne poses four internal problems faced by cartels. The "locating" (of the contract surface) and "sharing" problems together pertain to the determination of an optimal collusive policy; the "detecting" and "deterring" problems pertain to firms' incentives to unilaterally cheat on that policy. Claiming "Standard theory teaches us that cartels are inherently *unstable*, mainly because of the sharing and deterring problems" (p. 835), Osborne proceeds to argue that these are readily solved so that only the locating and detecting problems remain as obstacles to internal cartel stability.

Quite apart from his treatment and purported resolution of the sharing and deterring problems, which provide the focus for this comment, there is some question as to whether or not Osborne correctly interprets "standard theory." It would appear he does not. In particular his argument that economic orthodoxy stresses the deterring problem is misplaced; in fact it is the detecting problem that is generally emphasized. For example in his treatise on monopoly, Donald Dewey (p. 19) indicates the importance to a cartel's success of reliable information on the behavior of member firms. The information is necessary, Dewey claims, to detect cheaters. In a recent volume on antitrust law, Richard Posner writes, "Cheating is presumably least likely when *detection* is prompt and certain..." (p. 53, emphasis added). George Stigler certainly emphasizes the detecting problem over the problem of deterrence; indeed he does not view deterring per se as a problem: "Once detected, the deviations [i.e., cheating] will tend to disappear because they are no longer secret and will be matched by fellow conspirators if they are not withdrawn" (p. 42). Osborne's exegesis of standard theory is incomplete. Detection is a serious and widely recognized problem that has been emphasized more than deterrence.[1]

Questions of economic literature aside, even if deterring is to be granted equal status with detecting as a cartel "problem," its importance depends crucially on the extent of the detecting problem. As we shall show, Osborne's purported resolution of the deterring problem is either ineffectual or unnecessary depending on the status of the detecting problem. If the cartel has a detecting problem, then his resolution of the deterring problem breaks down. If detecting is not a problem for the cartel, then following Stigler, neither is deterring, since detection alone is sufficient to bring an end to cheating. Moreover if deterring persists as a problem, as Osborne seems to aver, even when there is no detecting problem, his resolution of it is not strictly superior to an infinite number of alternative resolutions.

We also argue that Osborne's resolution of the sharing problem, although correct in the case of a single joint profit-maximizing point, is unconvincing for the case of multiple joint profit-maximizing points. Finally, in spite of its failure as a final and complete resolution of the deterring problem, Osborne's "quota rule" has some interesting implications for still another (domestic) cartel problem: the avoidance of antitrust prosecution and conviction.

*Assistant professor of economics and professor of economics, respectively, University of Virginia.

[1]In the profession's best-selling principles text, presumably an indicator of standard theory, the detection of secret price concessions is stressed, not their deterrence (see Campbell McConnell, p. 581) The only source found that placed a stress on deterrence over detection was Gary Becker, pp. 100-01. F. M. Scherer, in his discussion of collusion, draws no clear distinction between the detection and deterrence aspects of cheating, pp. 158-64. In his later examination of trade associations, however, he does cite the deterrence aspect, p. 449.

FIGURE 1

I. Deterring

To illustrate the objection to Osborne's resolution of the deterring problem, consider the two-member cartel case[2] shown in Figure 1: (X_1^0, X_2^0) is the "cartel point," maximizing the members' joint profits over the contract curve (and minimizing the variance of members' profits over the set of joint profit-maximizing points). Osborne shows that this point possesses the "ray property": the ray passing through it is tangent to both members' iso-profit curves I_1^0, I_2^0 at the point. The collusive agreement that allegedly solves the problem of deterring is the following quota rule for member i:

$$\text{Produce} \qquad \max\left\{X_i^0, X_i^0 + \frac{S_i}{S_j}\,\Delta X_j\right\}$$

where ΔX_j is the amount by which member j's output deviates from his quota X_j^0, and where $S_i = X_i^0/\Sigma_j X_j^0$. This agreement specifies two things for each member: his output level at the cartel point; and the retaliatory rule he is to follow in the event another member cheats by increasing output above an alloted level.

The appeal of this rule as a deterrent is because "member j will not cheat if he expects the other members to obey their quota rules." If member 1 cheats by increasing his output from X_1^0 to X_1^1, he expects that member 2 will retaliate by increasing his output from X_2^0 to X_2^1. This penalizes rather than rewards the former for his recalcitrance. And member 1 is said to have every reason to expect member 2 to retaliate in this way because his profits at (X_1^1, X_2^1) are greater than at (X_1^1, X_2^0). Indeed, such an agreement would seem to deter each member from deviating from the cartel point. But it may not.

Suppose there is a detecting problem—that is, assume that immediate detection is not certain. Deviating from the cartel point in this case may be quite rational. Depending on the probability distribution of time until detection occurs, the expected gains from increasing output above the level alloted at the cartel point may be positive. If such an incentive exists for a member to cheat on the cartel point however, the same incentive exists for another to cheat on the quota rule retaliatory mechanism. If instantaneous detection is not certain, member 2 might increase his output from X_2^0 to X_2^2 rather than merely from X_2^0 to X_2^1. The expected gains from this second kind of cheating may be no less substantial than those for the first. It is a strange thing that cartel members who are *not* trusted to maintain the cartel point aspect of an agreement *are* trusted to maintain the quota rule aspect. There is, then, no reason to suppose that the quota rule successfully deters when there is a detecting problem. Osborne's alleged resolution is ineffectual here.

Now suppose there is *no* detecting problem—that is, assume immediate detection is certain. In what sense is it now rational for members to deviate from the cartel point? As mentioned earlier, Stigler suggests that cartel members implicitly assume that punitive[3] retaliation will follow detection. If

[2]We implicitly adopt the technological assumptions stated by Osborne.

[3]Retaliation is punitive if it makes the cheater worse off after detection and retaliation than before cheating.

this is correct, then there is no deterring problem to be resolved. If, however, members don't operate under such an assumption, *any* retaliatory agreement that is well understood and punitive will solve the deterring problem. Osborne's is only one among many. It has the virtue that it is easily articulated and executed, but it is not uniquely effectual. It is, as Osborne indicates, the "cheapest deterrent": of all the retaliatory rules that are punitive, it displaces the cartel less from the cartel point than others. But whether or not this is a virtue is questionable. If a retaliatory rule is successful—as many in this case should be—then it would never be exercised. So what does it matter that it is "cheaper" than others *if it were* exercised? It won't be. The quota rule, then, is no more or less effectual in this case than many other rules. Osborne's alleged resolution is unnecessary here.

II. Sharing

It is also alleged that the quota rule helps solve the sharing problem. In the case of a single joint profit-maximizing point, the sharing problem should be resolved, and for the reason Osborne offers. Any other point will not prevail because of the arbitrage incentive any member would have to purchase control of the others, move to the cartel point, and retain the difference in profits. His "deeper reason why the quota rule solves the sharing problem" is that it solves the deterring problem and therefore makes the cartel point a safer solution to the problem of sharing. The points in the previous section depreciate the persuasiveness of this reason, although the first reason is sufficient alone to solve the sharing problem.

In the case of multiple joint profit-maximizing points, Osborne appeals to the notion of a focal point in claiming that cartelists will opt for that point which minimizes the variance of profits. This is less than persuasive. Focal points of coordination are drawn from particular institutional settings and circumstances and as such will vary as between industries. In his seminal discussion, Thomas Schelling argues that a focal point "... may depend on imagination more than logic; it may depend on analogy, precedent, accidental arrangement, symmetry, aesthetic or geometric configuration, casuistic reasoning, and who the parties are and what they know about each other" (p. 57). A price-fixing cartel of independent gasoline marketers, for instance, might set their joint price at a constant differential from the current retail price of the major producers.[4] No other industry would settle on such a focal point. In a market-sharing cartel of two firms, where each firm had all its production facilities on opposite sides of the Mason-Dixon line, that fact of geography might provide the focus for market division. In the case at hand, the member's precartel market status might provide the focus for choosing from among several joint profit-maximizing points. The point is that different cartels are likely to have different focal points and what makes them attractive is not generally the minimum variance property Osborne suggests.

III. Antitrust Implications

Finally, there are some antitrust implications of the quota rule for cartelists operating in the United States. The Sherman Antitrust Act makes market-sharing agreements per se illegal. Given the costs imposed upon convicted colluders, in the form of fines, incarceration, and treble damage suits, antitrust can be viewed as another external problem to cartel stability.[5] Rational car-

[4]This is indeed one of the tactics used by a regional cartel of gasoline marketers recently convicted of price fixing See "Five Oil Firms, Group Convicted In Price-Fix Plot," *Wall Street Journal*

[5]In a footnote added in proof, Osborne argues that the one "fatal problem" to a cartel is the external threat from new entry or new products. He states he could find no record of a "cartel which died of internal problems alone." Without gainsaying the general validity of his conclusion, the plumbing fixtures cartel does provide a contrary example. While plagued by the external problem from time to time, it nevertheless met its demise because of problems other than new entry or new products. In this remarkable conspiracy, W. Kramer, executive secretary of the Plumbing Fixtures Manufacturers Association (*PFMA*), had the

telists will consider the effects of their actions upon the probability of detection and conviction under the antitrust laws (see Elzinga and William Breit, pp. 116–20). In the situation where instantaneous detection is not certain, the quota rule has characteristics not possessed by other retaliatory rules which both cut for and against the avoidance of antitrust prosecution. Its usefulness in eschewing the law stems from its simplicity of operation. Because it can be followed independently by each member firm, the need for direct meetings or communication is minimized, thereby reducing the tangible evidence of collusion. Adherence to the rule leaves no trail of hotel meetings or telegrams. Its antitrust drawback is that following the rule leaves member firms with their original market shares. If the antitrust authorities are watching market share stability as a sign of collusion, as has been proposed (see, for example, Mark J. Green, p. 177), the quota rule provides a tip-off as to the violation.

cartel sessions secretly taped. Then he embezzled funds from the *PFMA*, purchased a yacht, and left the country When the *PFMA* demanded his return with the funds, he blackmailed the members with the tapes The cartel ended The conspiracy became public when the Internal Revenue Service, on Kramer's trail for tax evasion, uncovered the incriminating tapes (see Allan T Demaree) At that point, the other external threat to cartels, antitrust prosecution, became a reality

REFERENCES

Gary S. Becker, *Economic Theory*, New York 1971.

A. T. Demaree, "How Judgment Came for the Plumbing Conspirators," *Fortune*, Dec. 1969, *80*, 96ff.

Donald Dewey, *Monopoly in Economics and Law*, Chicago 1959.

Kenneth G. Elzinga and William Breit, *The Antitrust Penalties: A Study in Law and Economics*, New Haven 1976.

Mark J. Green, *The Nader Report: The Closed Enterprise System*, New York 1972.

Campbell R. McConnell, *Economics*, New York 1975.

D. K. Osborne, "Cartel Problems," *Amer. Econ. Rev.*, Dec. 1976, *66*, 835–44.

Richard Posner, *Antitrust Law*, Chicago 1976.

Thomas C. Schelling, *The Strategy of Conflict*, Cambridge 1960.

F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

G. J. Stigler, "A Theory of Oligopoly," *J. Polit. Econ.*, Feb. 1964, *72*, 44–61.

Wall Street Journal, Aug. 31, 1977, p. 2.

Cartel Problems: Comment

By WILLIAM L. HOLAHAN*

Cartels are inherently unstable. At the joint profit-maximizing price and output every member has an individual incentive to expand output, secretly if possible, and cheat on the cartel even though it is better off with the cartel intact than if the cartel dissolved and the members competed. Dale Osborne has recently proposed a rule for cartel members which, if followed, will pose a credible threat of lost profits to potential cheating members and thereby reduce the inherent instability of the cartel. The rule is simple: once cheating in the form of increased output is detected, each member should increase output in the same proportion as the cheater so as to maintain the same share of the output as under joint profit maximization. This market share maintenance rule forces the cheater to share in the decline of profits and hence induces it to help the loyal members revitalize the cartel or, perhaps, not cheat in the first place. Following this rule the cartel should be far more stable than traditional theory would predict, consistent with the recent history of the Organization of Petroleum Exporting Countries (*OPEC*) which has remained remarkably stable despite prices incredibly far above some members' costs.

The purpose of this comment is to point out some improvements in Osborne's analysis. In Section I it is shown that his proof that the market share maintenance line has a common tangency with all of the cartel members' iso-profit surfaces at the point of joint profit maximization is too restrictive. A more general proof is provided. In Section II it is pointed out that his proof that the market share maintenance rule provides the noncheater a profit-increasing retaliation against the cheater is not valid, but that the rule retains many advantages which Osborne does not mention. Section III points out some advantages of central purchasing agencies which Osborne has overlooked in his section on purchasing strategies.

*Assistant professor of economics, University of Wisconsin-Milwaukee. I wish to thank John P. Brown of the Cornell economics department and Robert Hall of the University of Wisconsin-Milwaukee mathematics department for their helpful comments on an earlier draft.

I

Osborne assumes (Assumption 2, p. 837) that the cross-income effects are either equal or negligible in order to complete his proof that the market share maintenance line has a common tangency with all of the iso-profit surfaces at the joint profit-maximizing point. A more general proof without this restrictive assumption is as follows.

Let there be n cartel members. The ith member produces x_i $(i = 1, \ldots, n)$. The profit function of the ith member is:

$$(1) \qquad \Pi_i = x_i P(\Sigma x_i) - C_i(x_i)$$

The objective of the cartel is:

$$(2) \qquad \max_{x_1, \ldots, x_n} \sum_j \left[x_j P(\Sigma x_i) - C_j(x_j)\right]$$

The n first-order conditions for this maximization are

$$(3) \qquad P + P' \cdot (\Sigma x_i) - C_j' = 0 \qquad j = 1, \ldots, n$$

At the joint profit-maximizing point $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$, described by (3), each iso-profit surface is tangent to a hyperplane. Osborne labels these hyperplanes $H_1, \ldots, H_n$. Hyperplane H_j can be described by

$$(4) \qquad \sum_{i=1}^{n} \alpha_{ij}(x_i - \bar{x}_i) = 0$$

where the α_{ij} are constants. Arbitrarily selecting x_k for normalization yields

$$(5) \qquad x_k - \sum_{i \neq k} \beta^j_{ik} x_i = \psi_{kj}$$

FIGURE 1

where ψ_{kj} is the intercept of H_j along the x_k axis and $\beta^j_{ik} > 0$, $i \neq k$, are the slopes of H_j in the x_k, x_i plane. Using the implicit function theorem on the jth iso-profit surface $\Pi_j = \Pi_j(\overline{x})$, we can write

$$(6) \quad \beta^j_{ik} = \frac{dx_k}{dx_i} = -\frac{\partial \Pi_j / \partial x_i}{\partial \Pi_j / \partial x_k}$$

$$= \begin{cases} -1 & i \neq j, i \neq k \\ -\dfrac{x_j P' + P - C_j'}{x_j P'} & i = j, i \neq k \end{cases}$$

Substituting the $n - 1$ expressions given by (6) into (5) and simplifying yields:

$$(7) \qquad \frac{(P + P'(\Sigma x_i) - C_j')}{P'} = \psi_{kj}$$

But the parenthetical expression in the numerator of (7) is just one of the first-order conditions given by (3). Hence $\psi_{kj} = 0$ for all k, H_j intersects each axis at the origin. This is true for all j. Hence each hyperplane has the origin and $\overline{x}$ in common. Since all the hyperplanes have two points in common, their intersection is a line. Since the hyperplanes have the origin in common, the line has the ray property—the line is a market share maintenance line. Osborne's restriction on the cross-income effects is not required for this result.

II. The Market Share Maintenance Rule

Consider Figure 1 which illustrates in two dimensions the cartel stability problem and the market share maintenance rule. The axes depict the output of the cheating coalition on the abscissa and the remainder of the cartel, the loyal coalition, on the ordinate. The curved lines are iso-profit lines (subscripted with l for loyal and c for cheater) drawn concave to the respective axes. Joint profits are maximized at point a where iso-profit lines π_l^a and π_c^a are tangent to line L, which has been shown in Section I to be a ray from the origin. Point a is unique since there is only one allocation which simultaneously equates marginal revenue to marginal costs for all members. Since L is a ray from the origin it is a locus

of points in output space such that market shares are constant. The market share maintenance rule in graphical terms is: if cheaters expand output along line C, retaliate to a point along line L. For example if the cheater moves to b, retaliate by a move to d vertically above b on L. Due to concavity of π_c^a, $\pi_c^d < \pi_c^a$; the cheater is worse off after the retaliation. Osborne refers to this retaliation as a "credible" threat because, again by concavity, $\pi_l^d > \pi_l^b$; the loyal coalition is better off if it retaliates to d than if it remains at b. Such retaliation will deter a move to b. This is the essence of Osborne's proof of the credibility of the market share maintenance rule.

The difficulty is that the proof that the loyalists can enjoy profit-increasing retaliations by employing the market share maintenance rule does not necessarily hold throughout the plane, and hence retaliation along L is not credible according to Osborne's usage of that term. Figure 1 is drawn to point out an important counterexample in which the cheaters could move to the profit-maximizing point g instead of b. In fact line R_1 is the locus of iso-profit curve maxima and hence is the locus of the cheaters' profit-maximizing (Cournot) reactions to any output set by the loyal members. Similarly, R_2 is the profit-maximizing reaction function of the loyal coalition. The profit-maximizing reaction of the loyal coalition to a move to g is not to follow the market share maintenance rule to h, but instead to act as a Cournot reactor and move to i. The cheaters would then move to j; iterations of this kind lead to the Cournot solution at c. At c: $\pi_c^c > \pi_c^a$, $\pi_l^c < \pi_l^a$; the cheating has paid off in increased profits to the cheaters, less profits for the loyal coalition and less in total. An even more profitable form of cheating is to move along line C to a point to the right of m (which is vertically above point c) such as k. Then the iterative reactions lead to s, the Stackelberg solution. Here again the market share maintenance rule is not credible by Osborne's definition because such retaliations make the retaliators even worse off than standing pat; the profit-maximizing reaction is in the other direction.

This result was obtained because the reaction functions crossed inside the iso-profit line π_c^a. This could not happen if the two families of iso-profit lines were symmetric; in that case the market share maintenance rule would be the 45° line and by symmetry the reaction functions would cross on 45° line. The market share maintaining retaliations would be profit increasing for the loyalists in the symmetric case. The cheater would be worse off and induced to revitalize the cartel at the joint profit-maximizing point.

Figure 1 is drawn purposely to represent a nonsymmetric case in which the cheating coalition's iso-profit curves have less curvature than the loyal coalition and the reaction functions cross inside the cheater's π_c^a iso-profit curve. This can arise due to either lower costs or a higher perceived elasticity of demand due to a longer time horizon. The iso-profit lines for the coalition with the longer time horizon will have less curvature than the lines for the coalition with smaller reserves and shorter time horizon—for any given output by the loyal members, the cheater's profit-maximizing output is larger as the time horizon is longer. Equivalently, the longer the time horizon the greater the incentive to induce a Cournot or Stackelberg leadership position.

It is essential to realize that the cheater cannot know in advance that conditions are favorable. It can suspect that $\pi_c^c > \pi_c^a$, as drawn in Figure 1. But it may fear that instead the reaction curves cross above the π_c^a iso-profit curve and that $\pi_c^c < \pi_c^a$ and $\pi_c^s < \pi_c^a$; the Cournot iterations or the Stackelberg equilibrium lead to less profits. The cheater can at most know his own iso-profit functions. Since the cheater does not know the loyalists' iso-profit functions it cannot locate points c and s a priori. The cheater can only find these points by cheating and observing the loyalists' response. If the loyal coalition employs the market share maintenance rule the cheater will have a smaller net revenue flow than at a

and may be induced to return to a. If the loyalists fall into line at point c or s, the cheater has a more permanent gain. The cheater can increase the probability of a profitable result by announcing its action. It wants to become a Stackelberg leader or at least induce a Cournot equilibrium. It can only do so if the loyal coalition remains organized enough to follow its lead. If it didn't announce, the loyal coalition would not know the source of the extra output. Distrust would grow among the loyal coalition and it would weaken. By the cheater's announcement, the source of the extra output is identified and this reinforces commonality of interests among this subcartel. This reduces the probability of a collapse to competitive equilibrium after the cheater cheats.

At the Cournot or Stackelberg equilibrium there is the possibility of mutual gain from payments from the loyal members in exchange for a reformation of joint profit maximization. For example if the equilibrium is at s, a payment B from the loyal members to the cheater, such that $B > \pi_c^s - \pi_c^a$, would induce the cheater to return to point a. This payment can be made by the loyalists if the payment is less than the gain: $\pi_c^s - \pi_c^a < \pi_l^a - \pi_l^s$. But this is equivalent to $\pi_c^a + \pi_l^a > \pi_c^s + \pi_l^s$ which is always true since joint profits are maximized at a. These side payments can help stabilize the cartel since they would cease when cheating started.

It is in the interest of the loyalists to adopt strategies which avoid the Cournot or Stackelberg equilibrium points. But here we have seen that profit-increasing retaliations can lead toward such equilibria. The loyalists may adopt strategies which decrease their own short-run profits as well as the cheater's in order to induce the cheater to return to the joint profit-maximizing point. This strategy may increase the present value of the loyalists' assets in spite of short-run losses. Such strategies are credible responses in a broader usage of that term.

With such strategies the loyalists and the cheater play a waiting game under uncertainty. Neither the cheater nor the loyalists knows the other's iso-profit curves or reaction function. The cheater attempts to ferret out the loyalist's reaction function by cheating. If the relative positions of the reaction functions are favorable to the cheater, as shown in Figure 1, the loyalists must act to hide this information by *not* selecting the profit-maximizing reaction. If they are to convince the cheater to rejoin the cartel they must adopt a short-run strategy which will definitely hurt the cheater and which will make the cheater think that perhaps the loyalists' retaliation is profit increasing. The market share maintenance rule has some advantages for the loyalists in this game: 1) it is a simple way to coordinate the retaliation; 2) it hides iso-profit information from the cheater; 3) it gives the loyalists the appearance of orderly retaliation since the strategy was announced prior to the cheating; 4) it assures the loyalists that the cheater is definitely worse off; 5) of all the retaliations which give this assurance, it gives the loyalists the largest net revenue flow.

To see these results consider a move to k by the cheater. Retaliation to l reduces the loyalists' profits. The loyalists know that the cheater is worse off than at k but does not know if the cheater is worse off than at a. The cheater does not know if l lies on the loyalists' reaction function or if the loyalists are bluffing. A retaliation to n instead of l leaves the cheater equally in the dark but does assure the loyalists that the cheater is playing the waiting game with less net revenue than at a. Any other retaliation with this assurance would lie vertically above n and entail a smaller net revenue flow for the loyalists during the waiting game. Hence, the market share maintenance rule has cartel stabilizing features even though Osborne's proof is not valid.

III. Osborne on Purchasing Strategies

What strategies are left to consumers facing a cartel stabilized by the market share maintenance rule or one which has collapsed into a Cournot or Stackelberg

equilibrium? Since the various coalitions or subcartels have their own inherent instability problems one would expect that they could be weakened by proper purchasing strategies. Osborne argues in opposition to a central purchasing strategy because it would identify sellers and ease the cartel's (or subcartel's) most difficult problem—detection of cheaters. This would not be true if the bids were sealed and the bid prices and quantities kept secret by the central agency. The sovereign cartel members who wish to cheat will find an institution which provides the desired secrecy. This arrangement would encourage competition among the cartel members because it would create an uncertain climate conducive to breakdown. If the bidding is done by a common purchasing agency it will keep prices down for consuming nations. If it were not a monopsony, the producing nations would simply ignore it and sell to competitive bidders through a common sales agency. It is easier to hold a common sales agency together if there is competitive purchasing than if the cartel faces a monopsony with power to encourage secretive cheating.

REFERENCE

D. K. Osborne, "Cartel Problems," *Amer. Econ. Rev.*, Dec. 1976, *66*, 835–44.

Cartel Problems: Reply

By D. K. OSBORNE*

I. To David Mills and Kenneth Elzinga

I enjoyed the story of the plumbing fixtures cartel being drained of its assets. But if I need a "plumber's friend," David Mills and Kenneth Elzinga have failed me. With the possible exception of their second paragraph,[1] nothing in their comment convinces me to eliminate the offending material. I will try to explain why in paragraphs 1) and 2) below. Then I will comment briefly on the antitrust implications in paragraph 3).

1) Mills and Elzinga maintain that my resolution of the deterrence problem is either ineffectual or unnecessary. It is ineffectual in the absence of detection and unnecessary in its presence. They are right about the first. No deterrence is possible without detection. But who would think otherwise, or that I had claimed otherwise? As for the second, they appear to believe that deterrence follows immediately from detection. This belief is obviously mistaken—as our crowded jails prove. More direct proof is furnished by the experience of the International Air Transport Association (*IATA*). For detection, this cartel depends on our Civil Aeronautics Board and Department of Justice and its own compliance department (consisting of some fifty investigators) to inspect tickets, receipts, and accounting records at offices of the members and their approved travel agents (see *IATA Review*). For deterrence, it relies on the fines determined by due process before its Breaches Commission or the federal courts. In 1974, the Breaches Commission levied fines of $1.9 million (see *Aviation Week*); in fiscal 1975, the Civil Aeronautics Board obtained judgments totaling $556,594 (see its *Reports to Congress*); in September of 1975 the Justice Department obtained fines totaling $655,000 (see *Aviation Week*). These fines measure *IATA*'s success at detection and its failure at deterrence. In the same way that lax enforcement of the criminal laws leads to jails full of prisoners, inadequate punishment of cartel breaches increases their expected payoff and stimulates both the breaches and the fines in which they result. The *IATA*'s penalties have been too small or too uncertain to be regarded as anything more than a normal cost of doing business. Deterrence does not follow from detection.[2]

2) Mills and Elzinga object to the minimum-variance criterion for choosing among several joint maximizing points. I will be glad to consider an alternative if Mills and Elzinga will offer one. Instead of suggesting a definite alternative or, even better, the principles which govern its choice, they declare in effect that it could be anything.[3] If focal points really do depend on analogy, accident, casuistry, and the other things in the list quoted from Thomas Schelling, they are analytically useless: being consistent with everything they explain nothing.

3) Mills' and Elzinga's remarks about antitrust implications disquiet me. They themselves do not recommend that we go about prosecuting oligopolists for having stable market shares, but their remarks alert me to the danger that my analysis will

*Federal Reserve Bank of Dallas.

[1]There they object to my mild chastisement of standard theory for its fascination with the prisoners' dilemma. But since they advance the objection in an incidental manner I will disregard it, pointing out, however, that standard theory consists of the unwritten as well as the written word. For a more detailed criticism of the prisoners' dilemma as a model of oligopoly, see my (1976) paper.

[2]For more on these problems, see my (1977a,b) papers.

[3]Concerning the two examples they give (the gasoline marketers and the firms separated by the Mason-Dixon line), it is impossible to say whether the assumed arrangements indicate the unique joint maximum or a choice from among many.

be misused by those who lack a sense of the absurd. I therefore urge the following considerations on those who see collusion everywhere in the American economy.

The quota rule deterrent (as Mills and Elzinga undoubtedly realize) is objectionable on empirical grounds. There appears to be no evidence of a cartel relying on retaliation in kind, through the preservation of market share, to deter breaches. (The "fighting ships" of the various shipping cartels, and the "fighting trade" generally, are species of retaliation in kind, but not through the maintenance of market share.) The most popular deterrent has been the fine. When used, it must have appeared to be the best available. Though the quota rule deterrent is optimal in the right conditions, those conditions must be rare. We need, therefore, to know more about the relation between conditions and optimal deterrents. How good must information be in order for the quota rule deterrent to be optimal? Under what conditions is the fine optimal? Judging by the variety of cartels that have used it, the fine must be optimal under a broad range of conditions. Are these likely to include the conditions faced by *U.S.* oligopolies subject to the antitrust laws? If so, any stability we find in their market shares probably does not indicate adherence to the quota rule. We need to answer these questions before using my analysis to identify collusion.

II. To William Holahan

William Holahan raises three objections, the second of which, though of uncertain practical significance, is theoretically valid and interesting. The first and third are simply mistaken.

1) The first objection is that my Assumption 2 is unnecessarily strong. (Assumption 2: $\partial f_i(x)/\partial x_j = \partial f_j(x)/\partial x_i$ for $i, j = 1, \ldots, n$, where f_i is member i's inverse demand function, x_i is his output, and $x = (x_1, \ldots, x_n)$.) This assumption may well be stronger than it needs to be. But far from dispensing with it, Holahan strengthens it. His equation (1) implies $f_1(x) = \ldots = f_n(x)$, i.e., perfect substitution, from which Assumption 2 follows immediately. (The converse is false.) As perfect substitution is essential to Holahan's equations (3) and (7), his apparently more general proof is actually less general than mine.

But I like the form of Holahan's approach to the ray property and want to clarify it for those who might prefer it to mine. If we replace his single inverse demand function $P(x_1 + \ldots + x_n)$ by the n functions $f_i(x)$ and correct his equations (2) and (3) as required, we can follow him through equation (5) and apply the implicit function theorem to get, in place of his equation (6), the correct equation (6′):

$$(6') \quad \beta^j_{ik} = -\frac{\dfrac{\partial \Pi_j}{\partial x_i}}{\dfrac{\partial \Pi_j}{\partial x_k}} \equiv -\frac{\partial F_j(x)/\partial x_i}{\partial F_j(x)/\partial x_k}, i \neq k$$

Substituting for β^j_{ik} in Holahan's equation (5) and simplifying, we get

$$(7') \quad \psi_{kj} = \frac{\sum_{i=1}^{n} x_i \partial F_j(\bar{x})/\partial x_i}{\partial F_j(\bar{x})\partial x_k}, j = 1, \ldots, n$$

which vanishes for all j if and only if $\bar{x}$ has the ray property. Hence Holahan's approach is an alternative to the one described on page 838 of my article and it might be more transparent to some readers.

2) The second objection—a valid one—is that market share retaliation does not always pay: if a member exceeds his quota by a sufficiently large amount, the remaining members will lose *more* by retaliating in full than by standing pat. While I admitted this on page 839, I went on to claim that *partial* retaliation would always be better than none. In terms of Holahan's diagram, my claim asserted the existence of a point g' vertically above g such that loyalist profit is greater at g' than at g. This cannot be true in the situation diagrammed, where the profit functions differ so greatly. In this situation, retaliation is not so credible a threat. It punishes the breacher al-

right, but at an immediate cost to the retaliator. Market share retaliation does not necessarily deter large breaches.

While this might in principle explain why no known cartel has used the quota rule, in practice there must be more to it. Surely the profit functions are sufficiently similar in some cases; and even when not, the feasible breaches must in some cases be sufficiently small. If a member wishes to increase his output from *a* to *g* but must do it in small steps that pass through *b* and *e*, then, depending on his discounting of the future, the threat of retaliation to the initial breaches can deter him even in the situation diagrammed. Since so much depends on how quickly output can increase, sales contract periods and capacity constraints play an important role. Cartels differ so in these respects that their neglect of the quota rule must, in some cases at least, have other explanations (the most likely one being slow and uncertain detection). Clearly, we have a lot to learn about the quota rule, and Holahan's criticism raises questions that merit further study.

3) The third objection is that, contrary to my argument, a central purchasing agency would not solve the cartel's detection problem because it could require sealed bids and keep prices and sales secret. Here there is a tacit appeal to the "sainthood" model of public employment: The agency's employees would possess information of great value to the cartel but would withstand every attempt at bribery because of their moral scruples, which are those of a saint. I believe there is some evidence against this model—though I do not of course speak from personal experience.

REFERENCES

W. L. Holahan, "Cartel Problems: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 942–46.

D. E. Mills and K. G. Elzinga, "Cartel Problems: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 938–41.

D. K. Osborne, "Two Notes on Oligopoly," *Z. Nationalökon.*, No. 1–2, 1976, *36*, 61–72.

______, (1977a) "Prospects for the OPEC Cartel," *Fed. Reserve Bank Dallas Rev.*, Jan. 1977, 1–7.

______, (1977b) "Dealing with OPEC," *The Wharton Magazine*, Summer 1977, *1*, 40–45.

Aviation Week and Space Technology, Oct. 6, 1975, 26.

Civil Aeronautics Board, *Reports to Congress, Fiscal Year 1975*, Washington 1976, 55.

International Air Transport Association, *IATA Rev.*, Sept. 1975, *10*, 5.

International Trade, Factor-Market Distortions, and the Optimal Dynamic Subsidy: Comment

By JAMES CASSING AND JACK OCHS*

When terms of trade change, productivity gains may be secured with a proper reallocation of labor. But such gains cannot be achieved costlessly. During the adjustment process it is likely that some labor will be unemployed. This raises the important problem of finding the adjustment path that optimally balances the gains from transferring labor against the cost of such transfers.[1] In a recent article in this *Review*, Harvey Lapan analyzes this problem within the context of a set of labor markets with the following characteristics: (a) "Institutional factors" maintain an equality of real wage rates in both the expanding and contracting sectors throughout the adjustment period. (b) The desired labor force in the contracting sector is determined by a profit-maximization condition, and is therefore controllable by a wage subsidy. (c) Each sector is served by a large labor market. The rate of migration of labor from the market serving the contracting sector to the market serving the expanding sector is a function $\phi(u)$ of the unemployment rate in the market serving the contracting sector, with the properties: $\phi(0) = 0$, $\phi' > 0$, $\phi'' \leq 0$, and $\phi'(0)$ is finite. Since the wage subsidy controls the quantity of labor demanded, it also controls the rate of transfer of labor from the contracting to the expanding industry. Given these assumptions, Lapan shows that in general the optimal rate at which labor should be forced out of the contracting industry is not zero; nor is it optimal to fully reallocate labor so as to equalize value marginal products. Therefore, an optimal policy generally requires wage subsidies in his model.

In Lapan's formulation, voluntary quits play no role in the description of the uncontrolled adjustment process. This may leave the reader with the impression that the divergence between the uncontrolled path and the optimally controlled path is created solely by the existence of involuntary unemployment. However, this is not the case. In Lapan's model the labor market is assumed to be subject to congestion since the conditions on ϕ imply that an individual's expected time of passage from one sector to the other is an increasing function of the quit rate. One purpose of our paper is to show that such congestion is a sufficient condition for the existence of a divergence between the private and social costs of separation during the adjustment process. Therefore, the argument for a wage subsidy does not rest upon the assumption that separations are involuntary.

Lapan assumes that the labor market is subject to congestion such that the probability of an individual passing through the market (securing a job) in a given period of time, $\phi'(u)dt$, is negatively related to the number of people in the labor market. However, he gives no explanation why one should expect labor markets to be congested. In the model presented below we characterize a labor market in which congestion will occur in the search for jobs.

I. The Nonoptimality of an Atomistic Adjustment Path

Following Lapan, suppose that there are $\bar{L}$ workers in the economy and two industries, 1 and 2. Each industry is made up

*Assistant professor and associate professor of economics, respectively, University of Pittsburgh. This paper stems from research sponsored by the U.S. Department of Labor, Bureau of International Labor Affairs – RFP-TIE-OFER-4, Research on the Impact of Foreign Trade and Investment Policies on U.S. Labor Markets. The paper has benefited from the helpful comments of Harvey Lapan and Marina v. N. Whitman.

[1] In principle, of course, one might consider the optimal adjustment paths of various interacting productive factors including labor.

of a large number of identical firms. Each firm in industry 1 produces according to a production function

$$(1) \qquad X_i = X_i(L_i)$$

where $X_i' > 0 \qquad X_i'' < 0, \qquad i = 1, 2$

Suppose further that domestic producers trade at world prices over which they exercise no influence. Starting from an initial equilibrium distribution of workers relative to an international price vector $P^o = (P_1^o, P_2^o)$, let the domestic economy adjust to a new world price vector P^1 where, say, $P_2^1 > P_2^o$ and $P_1^1 < P_1^o$. The reduction in the price of industry 1's output relative to that of industry 2 must induce a flow of workers from 1 to 2. Those who leave industry 1 must pass through the labor market. In any unit interval of time t, all those who are in the labor market are unemployed. A fraction F of those who are in the market, R(t), do not leave the labor market with a job in industry 2 at the end of the interval. This fraction F depends upon R(t). Therefore, $F(R) \cdot R(t)$ is the number of persons without employment at the end of interval t who will also be unemployed in period t + 1.

The function $F(R)$ represents impedance in our model. Accordingly, we wish to assume that $F' > 0$. The fraction of searchers failing to secure positions in any period increases with the number of searchers, that is, congestion exists.[2]

In order to justify this assumption we provide the following characterization of the labor market. Suppose there are S firms in industry 2. At the beginning of any time interval t, each firm determines a target labor force that will maximize its profits. Given its current labor force, the calculation determines each firm's current number of vacancies $V_s(t)$, $s = 1, \ldots, S$. The total number of vacancies for interval t is, therefore, $\Sigma_{s=1}^{S} V_s(t) = V$. We assume that the total number of people looking for work, $R(t)$, is such that $R(t) \leq V(t)$. That is, it is logically possible for everyone in the labor market to find work. Nevertheless, we suppose that no one knows how many individuals will show up at a given firm's employment queue at the start of each time interval. Each individual thereby finds himself in a game situation. Now, if each individual adopts a strategy that maximizes the likelihood of his getting a job given that everyone else chooses the same strategy, then each will choose the queue to which he goes by a random process. Conceptually, each individual will draw from an urn in which the proportionate distribution of balls representing each queue is chosen to maximize the probability that everyone will get a job.

The actual distribution of applicants over queues is therefore the result of the product of R independent "experiments." The probability of a set of experiments producing a particular distribution of applicants over queues, $\bar{r} = (\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_S)$, is then given by:

$$(2) \quad P(r) = \frac{R!}{\bar{r}_1! \bar{r}_2!, \ldots, (R - \sum_{s=1}^{S-1} r_s)!} \cdot \pi_1^{\bar{r}_1} \pi_2^{\bar{r}_2}, \ldots, \pi_S^{(R - \Sigma_{s=1}^{S-1} r_s)}$$

where π_i is the appropriate probability weight given the ith queue, $i = 1, \ldots, S$.[3]

Since all searchers are identical, the expected unemployment rate for any subset, $H \subset R$, is equal to the expected unemployment rate for the entire population R. The expected unemployment rate of a small subset, H, depends upon both the distribution of vacancies and the distribution of the remaining $Y = R - H$ workers. Consider an arbitrary distribution $\bar{y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_S)$, of Y workers over the S queues. Suppose R and every component of the arbitrary vec-

[2] $F' > 0$ is equivalent to Lapan's assumption that $\phi'' < 0$. James Tobin has suggested that no normative significance should be attached to the "natural" (i.e., steady-state) rate of unemployment because labor markets are characterized by congestion. However, Dale Mortenson has shown within the context of a simple congestion model that the unsubsidized steady-state rate of unemployment does have optimality properties. Nevertheless, since Mortenson's analysis focuses only upon comparative steady states, his result is not inconsistent with atomistic behavior producing a nonoptimal adjustment path.

[3] This is just the multinominal p.d.f. of $S - 1$ random variables $r_1, \ldots, r_{S-1}$.

tor $\bar{y}$ were to increase by a proportionality factor, $\lambda > 1$. Assuming the V_i's remain constant, the result of this increase in R is to reduce the number of vacancies available to members of H at each employment queue. Some queues which previously had a surplus ($V_i - \bar{y}_i > 0$) will now have no vacancies left for members of H. On this account alone, members of H, who search these queues with the same expected frequency as do all members of R, will now experience an increase in their expected unemployment rate. Since the choice of $\bar{y}$ was arbitrary, the (unconditional) expected unemployment rate for the subset H increases as R increases. Because the expected unemployment rates of H and of R are equal, the above argument implies $F'(R) > 0$.

If individual firms have only a fixed number of vacancies as our model implies, then congestion is a real phenomenon in labor markets; the market will not clear instantaneously, nor is it optimal to push all labor which will eventually be transferred into the market at the same time. Furthermore, the adjustment path produced by a set of individual expected income-maximizing decisions does not corresond to the optimal control path because no individual takes into account the impact of his quit decision on the unemployment experience of others.[4]

To establish the divergence between private and social costs, consider first the following multistage programming problem, designed to maximize national income after the exogenous terms of trade change over the planning horizon of length N:

$$(3)\quad \max \sum_{t=1}^{N} [W_1(t)L_1(t) + W_2(t)L_2(t)]$$

where

$$(4)\quad W_1(t) = P_1^1 X_1'(L_1(t))$$

$$(5)\quad W_2(t) = P_2^1 X_2'(L_2(t))$$

$$(6)\quad L_1(t) = L_1(0) - \sum_{\tau=1}^{t} [R(\tau) - F(R(\tau - 1))\cdot R(\tau - 1)]$$

[4]This is strictly true, as Lapan has noted in a private communication, only if the planning horizon is sufficiently long.

$$(7)\quad L_2(t) = L_2(0) + \sum_{\tau=1}^{t-1} (1 - F(R(\tau)))\cdot R(\tau)$$

$$(8)\quad L_1(t) + L_2(t) + R(t) = \bar{L}$$

and the initial conditions are given by

$$(9)\quad W_1(0) = \overline{W}_1$$

$$(10)\quad W_2(0) = \overline{W}_2 > \overline{W}_1$$

$$(11)\quad R(0) = 0$$

and a terminal state[5]

$$(12)\quad R(N + 1) = \epsilon$$

The state variables in the problem are wage rates—$W_1(t)$, $W_2(t)$—and employed labor—$L_1(t)$, $L_2(t)$. The control variables are the individuals in the labor market in interval t, $R(t)$. The various conditions simply require that wages reflect value marginal products and that everyone is somewhere in each interval. Those individuals in the labor market in interval t are there either because they were in the market during interval $t - 1$ and did not get work or because they quit at time t. The objective is to choose a path $\{R(t)\}$ that maximizes national income over the planning horizon.

Using Bellman's Optimality Principle, the recursive equation for the optimal performance function for a problem of length N is developed using[6]

[5]The terminal state condition simply captures the aspect that if only a fraction of searchers secure employment in any time interval, then any finite time horizon multistage dynamic programming solution necessarily yields some terminal state unemployment. The time horizon, however, is arbitrary.

[6]That this expression is indeed correct is easily seen. The terms $L_2(N - \tau - 1) + (1 - F(R(N - \tau - 1)))$ $R(N - \tau - 1)$ represents everyone employed in industry 2 in interval $N - \tau - 1$ plus the addition to that industry's employment at the beginning of interval $N - \tau$. The term $L_1(N - \tau - 1)$ corresponds to employment in industry 1 in period $N - \tau - 1$. From this, we subtract the *net additions* to searchers—that is, total searchers in period $N - \tau$ less searchers in period $N - \tau - 1$ who failed to find employment. These terms—$L_1(N - \tau - 1) + F(R(N - \tau - 1))\ R(N - \tau - 1) - R(N - \tau)$—are weighted by the wage in industry 1 at the beginning of interval $N - \tau$.

$$(13)\quad J^*_{N-\tau}(R_{N-\tau}) = \max_{R_{N-\tau}} \{W_2(N-\tau) \cdot L_2(N-\tau) + W_1(N-\tau)\cdot L_1(N-\tau) + J^*_{N-\tau+1}(R_{N-\tau+1})\} = \max_{R_{N-\tau}} \{W_2(N-\tau) \cdot (1 - F(R(N-\tau-1))R(N-\tau-1)) + W_2(N-\tau)L_2(N-\tau-1) + W_1(N-\tau)[L_1(N-\tau-1) + F(R(N-\tau-1))R(N-\tau-1)] - W_1(N-\tau)\cdot R(N-\tau) + J^*_{N-\tau+1}\}$$

where $\tau = 0, \ldots, N-1$.

Given that an optimal path is to be followed from time $N-\tau+1$ on, $J^*_{N-\tau}$ defines the optimal decision rule for the interval $N-\tau$.

Now define:

$$(14)\quad E^*_{N-\tau+1} \equiv J^*_{N-\tau+1} - W_2(N-\tau+1) \cdot (1 - F(R(N-\tau)))R(N-\tau)$$

The necessary conditions for optimal performance during the interval $N-\tau$ imply that the optimal $R^*(N-\tau)$ depends on both $F(R)$ and $F'(R)$. Specifically, along the optimal path in interval $N-\tau$,

$$(15)\quad \frac{dJ^*_{N-\tau}}{dR_{N-\tau}} = -W_1(N-\tau) + W_2(N-\tau+1)[1 - F(R^*) - F'(R^*)R^*] + \frac{\partial E^*_{N-\tau+1}}{\partial R_{N-\tau}} = 0$$

Suppose now that the optimal path will be followed from interval $N-\tau+1$ on out. Given the same wage differential at time $N-\tau$ as would exist along the optimal path, what is the value of $R(N-\tau)$ that would be produced by atomistic decision making? In particular, is it R^*?

We assume that each individual chooses a plan of action over the planning horizon that will maximize his expected income. The momentary equilibrium in interval $N-\tau$ therefore requires that the expected income of a person who enters the market in that interval be equal to his expected income if he delays entry. Given our assumption that wages and unemployment rates will follow the optimal path from $N-\tau+1$ onward and that workers anticipate this path of wages, atomistic momentary equilibrium in the interval $N-\tau$ requires:

$$(16)\quad -W_1(N-\tau) + (1 - F(R)) \cdot W_2(N-\tau+1) + \Big\{(1 - F(R)) \cdot \sum_{N-\tau+2}^{N} W_2(t) - \sum_{N-\tau+1}^{N} W_1(t) + F(R)\cdot G\Big\} = 0$$

where G is the expected value of income of a person who first enters the market at time $N-\tau+1$ and N is the last interval of the planning horizon.[7] Intuitively, this expression simply balances the expected loss and gain from entering the labor market.

The expression

$$(17)\quad \beta = (1 - F(R)) \sum_{N-\tau+2}^{N} W_2(t) - \sum_{N-\tau+1}^{N} W_1(t) + F(R)\cdot G$$

is simply the expected earnings gain over the interval $[N-\tau+1, N]$ for a worker who enters the market at time $N-\tau$. Therefore, assuming that the adjustment process follows the optimal path, we know

$$(18)\quad \beta = \frac{\partial E^*_{N-\tau+1}}{\partial R_{N-\tau}}$$

However, if $F' > 0$, then for interval $N-\tau$

$$(19)\quad -W_1 + [1 - F(R)]W_2 > -W_1 + W_2[1 - F(R^*) - F'(R^*)R^*] = 0$$

That is, if the optimal path is followed from time $N-\tau+1$, the atomistic path in interval $N-\tau$ will have a larger number of

[7]The expected gains and losses must balance. In period $N-\tau$, the foregone wage in the contracting industry $-W_1$ must just equal the wage in the expanding industry weighted by the probability of securing that wage $(1 - F(R))W_2$. And, the lost wages over the subsequent periods $\Sigma^N_{N-\tau+1} W_1(t)$, must similarly be offset by the sum of the probability weighted gain of employment being secured in period $(N-\tau)$, $(1 - F(R))\ \Sigma^N_{N-\tau+2} W_2(t)$, and the probability weighted gain of employment being secured in some other period, $F(R)\cdot G$.

searchers in the labor market than will the optimal path. It follows that the uncontrolled unemployment rate is always greater than the optimally controlled rate at the beginning of the adjustment process.[8]

Intuitively, each atomistic decision maker in making his optimizing calculation ignores the reduced probability of securing a job which he imposes on other searchers. This appears in inequality (19) as the term $-W_2F'(R^*)R^*$. Of course, in the absence of such "congestion," the atomistic path is indeed optimal. (This is clear from inequality (19), letting $F'(R) = 0$.)[9]

In our model, it is the appearance of a wage differential that provides the incentive for individuals to move to the expanding industry. Our particular specification of the wage-setting process must be viewed in perspective. We believe that our assumption that wages and vacancies are momentarily rigid is reasonably descriptive of reality. One explanation for such rigidity is that the steady-state number of vacancies is finite for each firm. Given set-up costs that are an increasing function of the number of workers hired, it may be unprofitable for a firm to instantaneously lower wages and expand its labor force beyond that sustainable in the steady state simply because of the appearance of more qualified applicants at its employment windows than there are vacancies posted.

In another sense, we do not believe our particular specification of the wage-setting strategy during the adjustment process is descriptive of reality. Given costs of search, all firms in the economy may be in a position to exploit limited (and temporary) monopsony power with respect to labor. Such exploitation, if possible, would make wages deviate from value marginal products during the adjustment process.

Deviations of wage from value marginal product would complicate the optimality analysis. The atomistic path is dependent on posted wages. But the optimal path is dependent on value marginal products, not posted wages. Our argument above shows that if wages are anticipated and equal to value marginal product the atomistic path is not optimal. However, if the actual wage-setting behavior reflects exploitation of monopsony power, then no definitive comparison of the atomistic and optimal control paths can be made without a precise specification of how monopsonistic wage setting affects the wage differential between industries. Nevertheless, the analysis above does imply that at the beginning of the process the initial number of searchers determined by atomistic adjustment to a wage differential equal to the differential in value marginal product is greater than the optimal number. Therefore, unless the exercise of monopsony power by all firms reduces the initial wage differential, atomistic adjustment will remain nonoptimal.

II. Summary

Lapan has shown that if workers have no control over the separation decision then interference in the market is generally required to produce an optimal adjustment path. His argument leaves open the question of whether intervention is called for when separations are voluntary. We have shown that in a decentralized labor market the search strategies adopted by individuals can lead to congestion. If the labor market is subject to "congestion," then there is a divergence between the private and social costs of search. This divergence leads to a

[8]Note that if the same number of people are transferred over the planning horizon along both the atomistic path and the optimal control path then the atomistic R cannot always be above the optimal R.

[9]In our model the cost of search is income foregone as a result of unemployment. Undoubtedly, many individuals search while still employed. Such on-the-job search is also costly if search time has any opportunity cost. If all search was conducted while on the job then the atomistic momentary equilibrium condition (equation (16)) would have to be expressed as balancing the expected cost of search against the expected wage gain from search. The basic point made in the above analysis is, of course, unaffected by whether the momentary equilibrium is expressed as balancing the expected cost of unemployment or the expected cost of search against the expected wage gain. In either case, congestion makes atomistic decision making nonoptimal during the adjustment process.

divergence between the atomistic and optimally controlled adjustment paths.

REFERENCES

H. E. Lapan, "International Trade, Factor Market Distortions, and the Optimal Dynamic Subsidy," *Amer. Econ. Rev.*, June 1976, *66*, 335–46.

D. T. Mortensen, "Job Matching under Imperfect Information," mimeo., Northwestern Univ., Oct. 1973.

J. Tobin, "Inflation and Unemployment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 1–18.

International Trade, Factor-Market Distortions, and the Optimal Dynamic Subsidy: Reply

By HARVEY E. LAPAN*

James Cassing and Jack Ochs' comment is, I believe, a very interesting extension of the analysis of my paper. Their two basic results are: (i) that congestion will occur in the search for jobs; and (ii) that given costly labor mobility, private decisions regarding voluntary quits will yield a socially optimal adjustment path if individuals have perfect foresight and if there is no congestion (externality) in the search process. Thus, they argue that even if factor prices are not rigid, the presence of congestion in the search process implies private decisions will not be socially optimal, and therefore that a subsidy will be needed to support the optimal plan.

While I agree with the conclusions of the Cassing-Ochs paper, I disagree with the proof they present. In deriving the socially optimal plan, the authors state the objective is to choose $R^*(t)$ (the number of workers searching for jobs) "...to maximize national income ... over the planning horizon" (p. 952), yet the objective function chosen (equation (3)) reflects only wage income, and not national income. If maximization of national income is the objective, then I believe the objective function should be:

$$(1) \quad \max \sum_{t=1}^{N} [P_1^t X_1(L_1(t)) + P_2^t X_2(L_2(t))]$$

where the notation is the same as in their paper. Optimizing (1), subject to their equations (6)–(8) (and the definitional equations (4)–(5)), the optimality condition reduces to (15), as presented in their paper. However, if their objective function is used ((3) or (13)), the optimality condition will not, I believe, reduce to (15); the reason for this is that, in differentiating (13) with respect to $R_{N-\tau}$, Cassing-Ochs (implicitly) treat $W_1(N-\tau)$ and $W_2(N-\tau+1)$ as constants. But, an increase in $R_{N-\tau}$ reduces $L_1(N-\tau)$, and increases $L_2(N-\tau+1)$, which, from their (4) and (5), implies that $W_1(N-\tau)$ and $W_2(N-\tau+1)$ change as $R_{N-\tau}$ changes. If the objective is maximization of *wage income* (as implied by their choice of objective function), then terms reflecting the changes in the wage rate due to changes in the control should appear in their objective function. Consequently, I believe that the criterion they present for an optimal path (15) is inconsistent with the objective function ((3) or (13)) that they use. On the other hand, if the objective is maximization of national income, as depicted by my (1), then I believe (15) reflects the appropriate optimality conditions. Nevertheless, I should stress that I do agree with their qualitative conclusion that congestion in the labor market will lead private decisions to be socially inefficient.

Furthermore, the model presented in my paper can readily be interpreted to consider the social optimality of private actions; the control model in no way assumes factor prices are rigid. The key assumption is

$$(2) \quad \dot{L}_c = \phi(u)L_m; \phi(0) = 0, \quad \phi' > 0, \phi'' \leq 0$$

where $\dot{L}_c$ is the increase in employment in C (the sector in which labor's marginal value product is larger); L_m is the stock of potential workers in M ($L_c + L_m = L$, constant), and u is the unemployment rate in M.

While the discussion of the optimal subsidy in my paper presumes unemployment is involuntary, nothing precludes us from interpreting u as voluntary unemployment (this distinction is irrelevant for a centrally controlled solution). In terms of the Cassing-Ochs paper:

$$(3) \quad R(t) = uL_m$$

where R is voluntary unemployment. Thus,

*Associate professor of economics, Iowa State University.

(2) depicts the relationship between job hires in C and the number searching for employment there (R(t)). Moreover:

$$(4) \qquad (d\dot{L}_c/dR) = \phi'(u)$$

since L_m(t) is given at t. Therefore, $\phi'(u)dt$ represents the probability that an individual searching for work for a time (dt) will find employment in C; it corresponds to $(1 - F(R))$ in the Cassing-Ochs paper. The assumption that $\phi'(0)$ is finite merely implies that, if there is only one searcher, it takes a nonzero amount of time for him (her) to find a job—a not unreasonable assumption.[1] From (4):

$$(5) \qquad (d^2\dot{L}_c/dR^2) = [\phi''(u)/L_m] \leq 0$$

Since (5) reflects the change in the probability of finding a job as the number of people searching increases, $\phi'' \equiv 0$ corresponds to no congestion ($F'(R) = 0$), whereas $\phi''(u) < 0$ corresponds to congestion in the labor market ($F'(R) > 0$).

Private individuals, in deciding whether to quit work and search, compare the opportunity cost of search to the expected benefits. Letting V(t) represent expected (private) net benefits of search:[2]

$$(6) \quad V(t) = \Big[\phi'(u) \int_t^T (W_c(\theta) - W_m(\theta))e^{-r(\theta - t)}d\theta - W_m(t)\Big]dt$$

In (6) $W_m(t)dt$ is the opportunity cost of searching for a time interval dt, $\phi'(u)dt$ is the probability of finding a higher paying job, and the integral represents the net discounted value of the higher wage rate.[3] Of course, T reflects the end of the horizon for the prospective searcher. If V(t) is positive at $u = 0$, some search is worthwhile; otherwise, none will be undertaken.

The socially optimal plan is given in my earlier paper; from the maximum principle (my (11)):

$$(7) \quad q\phi'(u) - PF'_m(N_m) \leq 0;$$
$$u[q\phi'(u) - PF'_m] = 0; \qquad N_m \equiv L_m(1 - u)$$

where q—the costate variable—is the (current) social value of an increase in L_c(t). The differential equation for q is (my (16)):

$$(8) \quad \dot{q} = (r + \phi)q - (F'_c - P(1 - u)F'_m) = (r + \phi - u\phi')q - (F'_c - PF'_m)$$

Consider the term $(\phi - u\phi')$; along an optimal path, u^*(t) is given. Define

$$(9) \qquad \epsilon(t) \equiv \phi(u^*) - u^*\phi'(u^*)$$

Given that $\phi'' \leq 0$, $\phi(0) = 0$, and $\phi'(0)$ is finite, then $\epsilon(t) \geq 0$ everywhere. Moreover, $\epsilon(t) \equiv 0$ if either (i) $\phi'' \equiv 0$, or (ii) $u^*(t) \equiv 0$ for all t. Thus:

$$(10) \quad \epsilon(t) > 0 \quad \text{if, and only if, } \phi'' < 0 \quad \text{and } u^*(t) > 0$$

Integrating (8), using (9) and the transversality condition $q(T) = 0$ yields

$$(11) \quad q(t) = \int_t^T [e^{-r(\theta - t)}e^{-\int_t^\theta \epsilon d\lambda} \cdot (F'_c(\theta) - PF'_m(\theta))]d\theta$$

In (11), r is constant, but ϵ is understood to depend on time (if $\phi'' < 0$ and $u^*(t) > 0$). Given q(t), u^*(t), the optimal unemployment rate at t, is determined from (7).

In order to compare the socially optimal plan to atomistic decisions, we must specify how W_c and W_m are determined. If these parameters do not reflect the current marginal value product of employed workers,

[1] The assumption that labor mobility is costless means, in our context, that $\phi'(u)$ is infinite.

[2] Formally, the decision is not only whether to search for a job, but when. Define $\tilde{V}(t) = V(t)e^{-rt}$, so that $\tilde{V}$(t) is the discounted value of search at t. If $0 \leq \tilde{V}(t) < \tilde{V}(t + dt)$, then search at t is not desirable, even though it will eventually become so; i.e., $u(t) = 0$, but $u(\tau) > 0$, some $\tau > t$. However, for $\phi'' \leq 0$, $r \geq 0$, and $\tilde{V}(t + dt) \geq 0$, it is readily shown that $u(t) = 0$ implies $\tilde{V}(t) \geq \tilde{V}(t + dt)$. Consequently, not all search will be postponed: $u(\tau) > 0 \rightarrow u(t) > 0$ for all $t < \tau$. Similarly, if employment never falls to zero in M (as is guaranteed by the Inada derivative conditions), then a competitive solution with perfect foresight implies $\tilde{V}(t) \leq 0$. Therefore, for a competitive solution, $u(t) > 0$ implies $\tilde{V}(t) = 0$; and $u(t) = 0$ implies $\tilde{V}(\tau) \leq 0$ for all $\tau > t$. Throughout, we assume individuals are risk neutral.

[3] The integral should run from $(t + dt)$ to T, but for small dt, the difference is of the second order of smallness ($(dt)^2$). Note that our formulation of the problem permits discounting, whereas Cassing-Ochs consider only the case where the private and social discount rates are zero.

it is clear private actions will not be optimal. Thus, assume

$$(12)\quad W_c(t) = F_c'(L_c(t));\ W_m(t) = PF_m'(N_m(t))$$

where C is the numeraire. Using (12) and (11), (7) becomes:

$$(7')\quad \phi'(u)\int_t^T [e^{-r(\theta-t)}e^{-\int_t^\theta \epsilon d\lambda} \cdot (W_c(\theta) - W_m(\theta))d\theta] - W_m(t) \leq 0$$

Comparing (7′) to (6), the private decision rule, we see that the expressions differ only in the term involving ϵ; if $\epsilon(t) \equiv 0$, the two expressions coincide, assuming individuals have perfect foresight. If there is no congestion ($\phi'' \equiv 0 \equiv \epsilon(t)$), then private decisions are socially optimal. Moreover, in my earlier paper I showed that for $\phi'' \equiv 0$, $N_m(t)$ increases over time (for $u(t) > 0$). This implies that all separations occur initially. As time passes, some individuals find jobs in C, whereas others return to their "original" jobs in M.[4] The pool of searching workers declines over time (this is also true for $\phi'' < 0$).

If $\phi'' < 0$, but $u^*(t) \equiv 0$ (i.e., if wage differentials are small, relative to r, or the length of the plan is short), then $\epsilon(t) \equiv 0$ and private decisions will again be socially optimal. However, if $\phi'' < 0$, and $u^*(t) > 0$ for some t, then (6) and (7′) no longer coincide. The congestion—or externality—causes private decisions to be socially suboptimal. Comparing (6) and (7′) we see that, starting from the same initial allocation of labor, the *initial* unemployment rate under private actions $u^P(t)$ will exceed $u^*(t)$, as stated by Cassing-Ochs. Clearly, the *initial* private unemployment rate is higher because private decision makers ignore the congestion caused by additional entries into the pool of searchers; the optimal plan properly recognizes these congestion costs.

However, this does not imply that private decisions lead to unemployment rates that are *everywhere* higher than for the optimal path. The higher initial $u^P(t)$ implies that at any future time, more workers will be employed in C under the private solution than under the socially optimal plan ($L_c^P(t) > L_c^*(t)$). Moreover, since the decision rule for terminating search—or labor transfers—is the same for private decisions and socially optimal ones; and since the terminal period of full employment increases as L_c increases, it immediately follows that full employment is restored sooner under private actions than under the socially optimal plan. Consequently, with congestion, the initial private unemployment rate is higher than is socially optimal, but ultimately it falls below the unemployment rate along the optimum path.[5] Full employment is achieved sooner under private actions, more labor is transferred to the higher wage sector, and national income—during the latter stages of the plan—is larger under private actions. Of course, the discounted value of national income over the whole plan is less under private actions.

The principal point raised by Cassing-Ochs, I believe, is that even if factor prices are not rigid, private actions will not be socially optimal if congestion occurs in the search process. This point—with which I completely agree—follows directly from the externality generated by too many people searching for jobs at any one time. Private decisions will be socially optimal if: (i) no congestion occurs; (ii) individuals have perfect foresight; (iii) the private and social planning horizons and discount rates are equal; and (iv) wages adjust instantaneously. Should any of these conditions fail to hold, then some intervention is required to support the socially optimal plan.

[4]This assumes no search is necessary in M—workers, having been employed there, know where to look for work. For symmetry, one should assume it is necessary to search for jobs in M as well as C; this, in turn would discourage initial quits. However, neither my original paper—nor the Cassing-Ochs specification—incorporates this assumption.

[5]Of course, if an optimal plan were instituted at any future moment, given the labor allocation provided by private decisions, then the socially optimal unemployment rate would be lower. However, it makes more sense to contrast the time path generated by private decisions to that generated by a plan that is optimal over the whole planning period.

REFERENCES

Cassing and J. Ochs, "International Trade, Factor-Market Distortions, and the Optimal Dynamic Subsidy: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 950–55.

H. E. Lapan, "International Trade, Factor Market Distortions, and the Optimal Dynamic Subsidy," *Amer. Econ. Rev.*, June 1976, *66*, 335-46.

The Genetic Determination of Income: Comment

By ARTHUR S. GOLDBERGER*

In quantitative genetics, the heritability of a continuous trait denotes the proportion of its variance which is attributable to genetic differences. A classical method for assessing heritability contrasts the correlation of the trait observed across pairs of identical twins (= monozygotic twins = *MZ*s) with that observed across pairs of fraternal twins (= dizygotic twins = *DZ*s). In its simplest version, the twin method attributes the greater correlation of *MZ*s entirely to the perfect correlation of their genotypes. Since the genotypes of *DZ*s, like those of ordinary siblings, correlate only about 1/2, the very simplest twin method just doubles the difference between the two observed correlations to estimate heritability.

Economists may have become aware of the heritability concept, and of the twin method, via the great IQ debate, in particular via the books of Arthur Jensen (1972, 1973), Richard J. Herrnstein, and Christopher Jencks. A series of articles by Paul Taubman (1976a,b), and Jere Behrman and Taubman, has now brought heritability and twin methods into economics itself. Twin data on schooling, initial occupation, later occupation, and earnings lead to such inferences as "Genetics by itself accounts for roughly 30 to 40 percent of everything except initial occupation, where it accounts for 8 percent" (Behrman and Taubman, p. 438).

I believe that such conclusions are unwarranted and indeed that the entire effort is misguided. The plan of this paper is as follows. In Section I, Taubman's equations are spelled out. In Section II, his procedure for obtaining bounds on parameters is analyzed. In Section III the specification of his model is critically examined. Section IV argues that heritability is essentially irrelevant for economic policy. An Appendix calls attention to other problems in Taubman's articles.

I. The Estimating Equations

Suppose that the variable observed for an individual, his phenotype Y is determined as the sum of two unobserved variables, his genotype X and his environment U:

$$Y = X + U$$

The two unobserved variables are symmetrically defined constructs: Genotype X is the expected value of Y for persons having a given genetic constitution, taken across the full distribution of environmental conditions. Environment U is the expected value of Y for persons developing in a given environmental condition, taken across the full distribution of genetic constitutions. Across individuals, the phenotypic variance is

$$\sigma_Y^2 = \sigma_X^2 + \sigma_U^2 + 2\sigma_{XU}$$

Division by σ_Y^2 gives the variance decomposition:

$$(1) \qquad 1 = h^2 + e^2 + 2rhe$$

where $h^2 = \sigma_X^2/\sigma_Y^2$ = heritability
= ratio of genotypic variance to phenotypic variance

$e^2 = \sigma_U^2/\sigma_Y^2$ = environmentability
= ratio of environmental variance to phenotypic variance

$r = r_{xu} = \sigma_{XU}/(\sigma_X\sigma_U)$
= correlation of an individual's genotype with his environment

*Professor of economics, University of Wisconsin-Madison. Research support has been provided by grants from the National Science Foundation, the Institute for Research on Poverty, and the Center for Advanced Study in the Behavioral Sciences. I am grateful also to John Conlisk, John Geweke, Robin Hogarth, Richard Lewontin, Michael Olneck, and Sandra Scarr for instructive correspondence and discussions. Responsibility for judgments and errors is solely mine.

The individual is paired with his twin, whose variables are distinguished by a prime:

$$Y' = X' + U'$$

Across such pairs, the phenotypic covariance is

$$\sigma_{YY'} = \sigma_{XX'} + \sigma_{UU'} + 2\sigma_{XU'}$$

the symmetry of the situation making $\sigma_{X'U} = \sigma_{XU'}$, and $\sigma^2_{Y'} = \sigma^2_Y$, $\sigma^2_{X'} = \sigma^2_X$, $\sigma_{X'U'} = \sigma_{XU}$. Division by σ^2_Y gives the correlation decomposition:

$$r_{yy'} = r_{xx'}h^2 + r_{uu'}e^2 + 2\,r_{xu'}h\,e$$

where $r_{yy'} = \sigma_{YY'}/\sigma^2_Y$
= correlation of an individual's phenotype with his twin's phenotype

$r_{xx'} = \sigma_{XX'}/\sigma^2_X$
= correlation of an individual's genotype with his twin's genotype

$r_{uu'} = \sigma_{UU'}/\sigma^2_U$
= correlation of an individual's environment with his twin's environment

$r_{xu'} = \sigma_{XU'}/(\sigma_X\sigma_U)$
= correlation of an individual's genotype with his twin's environment

For identical twins, $X = X'$, so $r_{xx'} = 1$ and $r_{xu'} = r_{x'u'} = r_{xu} = r$. Further, for identical twins, denote $r_{uu'}$ by p^*, and the observed $r_{yy'}$ by c^*. Then

$$(2) \qquad c^* = h^2 + p^*e^2 + 2rhe$$

For fraternal twins, on the other hand, the situation is less clear cut. Denote $r_{xx'}$ by g, $r_{xu'}$ by s, $r_{uu'}$ by p, and the observed $r_{yy'}$ by c. Then

$$(3) \qquad c = gh^2 + pe^2 + 2she$$

Equations (1)–(3) express the three observable items $(1, c^*, c)$ in terms of seven unknown parameters $(h^2, e^2, r, p^*, p, g, s)$.[1] Taubman assumes that $g = 1/2$ (the value which follows from supposing that mating is random and that all gene effects are additive), that $p = p^*$ (environments are as similar for DZ pairs as for MZ pairs), and that $s = pr$ (that is, $r_{xu'\cdot u} = 0$, the correlation of one twin's genotype with his DZ twin's environment vanishes when the first twin's environment is partialled out). I will assess these assumptions in Section III. For now, let us accept them, and thus have:

$$(4) \qquad 1 = h^2 + e^2 + 2rhe$$

$$(5) \qquad c^* = h^2 + pe^2 + 2rhe$$

$$(6) \qquad c = h^2/2 + pe^2 + 2prhe$$

Equations (4)–(6) constitute Taubman's estimating equations. They express the three observable items $(1, c^*, c)$ in terms of four unknown parameters (h^2, e^2, r, p). It is understood that $0 \le h^2, e^2 \le 1$ and that the observed data satisfy $0 < c < c^* < 1$.

In the classical twin method, it is assumed that $r = 0$, whereupon the system has a unique solution, namely

$$h^2 = 2(c^* - c)$$
$$e^2 = 1 - 2(c^* - c)$$
$$p = (2c - c^*)/(1 - 2(c^* - c))$$

(Notice how the double-the-difference estimator for heritability arises.) But Taubman (1976a, p. 861) considers $r = 0$ to be inappropriate and hence does not impose it. Instead, he assumes that r and p are nonnegative, so that the system (4)–(6) is subject to the inequalities

$$(7) \qquad 0 \le r \le 1 \qquad 0 \le p \le 1$$
$$0 \le h^2 \le 1 \qquad 0 \le e^2 \le 1$$

For his twin data, he tabulates selected numerical combinations of h^2, e^2, r^2, and p which are compatible with (4)–(7), emphasizing the maximum and minimum attainable values of those parameters.

Thus, for schooling, Taubman (1976a, pp. 865–66) observes $c^* = .76$ and $c = .54$, and reports these as the extreme points:

$$h^2 = .23,\ e^2 = .51,\ r^2 = .14,\ p = .55$$
$$h^2 = .46,\ e^2 = .54,\ r^2 = 0,\ \ p = .5701$$

For *log* earnings, he observes $c^* = .54$ and $c = .30$, and reports these as the extreme

[1] My h, e, p correspond to Taubman's g, n, ρ, respectively.

points:

$$h^2 = .06,\ e^2 = .63,\ r^2 = .58,\ p = .28$$
$$h^2 = .50,\ e^2 = .50,\ r^2 = 0,\quad p = .086$$

II. The Bounding Procedure

The extraction of such remarkably precise information—for example, the narrow ranges for e^2, r^2, and p for schooling—from such limited data should give one pause. Common sense dictates that we examine the bounding procedure analytically before giving credence to the results it produces in a particular data set.

To do so it is convenient to introduce these transformations of the observed correlations:

$$m = 2(c^* - c),$$
$$n = 2(1 - c^*),$$
$$t = 2(1 - c) = m + n,$$
$$k = 8c(1 - c^*) = 2n(2 - t)$$

With $0 < c < c^* < 1$, these observed quantities satisfy $0 < m, n, t, k < 2$. Next, it is convenient to solve (4)–(6) down to

$$r = (1 - h^2 - e^2)/(2he) \tag{8}$$
$$p = 1 - n/(2e^2) \tag{9}$$
$$e^2 = n(1 - h^2)/(t - h^2) \tag{10}$$

Equation (10) is the locus of h^2, e^2 combinations that satisfy (4)–(6). Let $f(h^2)$ denote the function on the right-hand side of (10). Its derivative, namely,

$$f'(h^2) = (1 - t)n/(t - h^2)^2$$

is signed by $1 - t$: For given data, the solution value for e^2 is monotonic in h^2, rising, remaining constant, or falling as h^2 rises according as $t < 1$, $t = 1$, $t > 1$. The solution values in (8)–(9) are also monotonic: r moves inversely with h^2, while p moves directly with e^2.[2]

Under the constraints (7), extreme values of h^2 and e^2 can occur at:

$$e = 1 - h$$
$$e^2 = 1 - h^2$$
$$e^2 = n/2$$

which correspond to $r = 1$, $r = 0$, and $p = 0$, respectively; $p \leq 1$ is not binding. Inserting each of these in turn into (10) locates the points:

$$A:\ h^2 = a^2,\quad e^2 = (1 - a)^2$$
$$B:\ h^2 = m,\quad e^2 = 1 - m$$
$$C:\ h^2 = 2 - t,\ e^2 = n/2$$

where a denotes the root $(0 < a < 1)$ of

$$h^3 - h^2 - (t + n)h + m = 0$$

the cubic equation which arises when $e = 1 - h$ is inserted into (10). Which of these points are attainable, and which pair of them locate the extremes, will depend on the observed data. For each point on the locus (10) the r and p values can be obtained from (8)–(9).

Analysis of the domain $0 < c < c^* < 1$ indicates that five data constellations, or cases, should be distinguished. These are defined below, in terms of the data (c^* and c) and again in terms of our transforms (m, n, t, k):

CASE 1: $1/2 < c,\quad n < 1 - m,\quad t < 1$

CASE 2: $1/2 = c,\quad n = 1 - m,\quad t = 1$

CASE 3: $c^*/2 \leq c < 1/2,$
$(1 - m) < n \leq 2(1 - m),\quad t > 1$

CASE 4: $c < c^*/2 \leq (2c)^{1/2} - c,$
$2(1 - m) < n \leq 2(1 - m) + 2k^{1/2},\quad t > 1$

CASE 5: $(2c)^{1/2} - c < c^*/2,$
$2(1 - m) + 2k^{1/2} < n,\quad t > 1$

For each case in turn, we locate the extreme points, and provide a numerical example.

CASE 1: The constraints $r \leq 1$ and $r \geq 0$ are binding, points A and B locate the extremes. As we move along the locus from A to B, decreasing r from 1 to 0, h^2 rises and (since $t < 1$) so do e^2 and p. For our numerical example, take Taubman's schooling data: $c^* = .76$, $c = .54$. Here $m = .44$, $n = .48$, $t = .92$. Thus the bounds are given by

[2]For details of the analysis and subsequent discussion see the more complete version of this article (1977b).

A: $h^2 = .076$, $e^2 = .525$, $r = 1$, $p = .543$
B: $h^2 = .44$, $e^2 = .56$, $r = 0$, $p = .571$

These differ from Taubman's tabulated extrema for schooling, cited earlier, so that his figures are incorrect. The discrepancies are quite minor, except for the upper limit on r, which clearly must be 1 rather than $.37 = (.14)^{1/2}$, and the lower limit on h^2 which should be .076 rather than .23.

CASE 2: The constraints $r \leq 1$ and $r \geq 0$ are binding, points A and B locate the extremes. As we move along the locus from A to B, decreasing r from 1 to 0, h^2 rises, but e^2 remains constant at $n/2$ (since $t = 1$) and p remains constant at 1/2. Note this singular case arises whenever the observed DZ correlation c happens to equal 1/2. For our numerical example, displace Taubman's schooling data slightly, taking $c^* = .72$, $c = .50$. Here $m = .44$, $n = .56$, $t = 1$. The bounds are given by

A: $h^2 = .063$, $e^2 = .52$, $r = 1$, $p = .50$
B: $h^2 = .44$, $e^2 = .56$, $r = 0$, $p = .50$

CASE 3: Again the constraints $r \leq 1$ and $r \geq 0$ are binding and points A and B locate the extremes. As we move along the locus from A to B, decreasing r from 1 to 0, h^2 rises, but now e^2 and p fall (since $t > 1$). For our numerical example, take Taubman's *log* earnings data: $c^* = .54$, $c = .30$. Here $m = .48$, $n = .92$, $t = 1.40$. The bounds are given by

A: $h^2 = .038$, $e^2 = .650$, $r = 1$, $p = .292$
B: $h^2 = .48$, $e^2 = .52$, $r = 0$, $p = .115$

These differ from Taubman's tabulated extrema for *log* earnings, so that his figures are incorrect. The discrepancies are minor, except for the upper limit on r which clearly must be 1 rather $.76 = (.58)^{1/2}$.

CASE 4: Here the constraints $r \leq 1$ and $p \geq 0$ are binding; points A and C locate the extremes. As we move along the locus from A towards B, decreasing r from 1 towards 0, h^2 rises, while e^2 and p fall (since $t > 1$). But we reach C ($p = 0$, $e^2 = n/2$) before arriving at B ($r = 0$, $e^2 = 1 - m$), since $n > 2(1 - m)$. For a numerical example, we take $c^* = .7$ and $c = .3$. Then $m = .8$, $n = .6$, $t = 1.4$, $k = .72$. The bounds are given by

A: $h^2 = .129$, $e^2 = .411$, $r = 1$, $p = .270$
C: $h^2 = .6$, $e^2 = .3$, $r = .117$, $p = 0$

CASE 5: Here no solution is admissible, the constraints are in conflict. At point A, p is already negative and falls as we move towards B. A numerical example is provided by $c^* = .9$ and $c = .2$, which give $m = 1.4$, $n = .2$, $t = 1.6$, and thus

A: $h^2 = .485$, $e^2 = .092$, $r = 1$, $p = -.087$

To supplement these examples, Tables 1 and 2 give for selected combinations of c^* and c, the upper and lower bounds on h^2 and e^2 which flow from Taubman's bounding procedure. Some features of the bounding procedure are now apparent:

(i) When $c = .50$, the upper and lower bounds on e^2 coincide: we are certain of the value of environmentability. (We are also certain that $p = .50$.) When c is in the neighborhood of .50, the upper and lower bounds on e^2 are quite close together. But if the data change slightly the bounds may move substantially. For example, if $c^* = .60$ and $c = .40$, we are certain that $.60 \leq e^2 \leq .66$, but if $c^* = .70$ and $c = .40$, then it is equally certain that $.40 \leq e^2 \leq .49$.

(ii) A slight change in the data can convert a narrow-bound situation into a no-solution situation. For example if $c^* = .60$ and $c = .10$, we are certain that $.12 \leq h^2 \leq .20$ *and* that $.40 \leq e^2 \leq .42$. But if $c^* = .70$ and $c = .10$, we are equally certain that the model does not fit at all.

(iii) The bounds are not confidence limits for they take no account of sampling variability. One might develop confidence limits on the bounds by perturbing the observed correlations, say, by ± one standard error as Taubman (1976a, p. 867) does. Alternatively, one can compute standard errors analytically, since the bounds are functions of the observed correlations.

TABLE 1—UPPER AND LOWER BOUNDS ON HERITABILITY h^2

MZ Twin Correlation c^*	*DZ* Twin Correlation c							
	.1	.2	.3	.4	.5	.6	.7	.8
.9	–	–	.60	.80	.80	.60	.40	.20
	–	–	.43	.37	.31	.23	.15	.06
.8	–	.40	.60	.80	.60	.40	.20	
	–	.28	.24	.19	.14	.08	.03	
.7	–	.40	.60	.60	.40	.20		
	–	.17	.13	.09	.05	.02		
.6	.20	.40	.60	.40	.20			
	.12	.09	.06	.04	.01			
.5	.20	.40	.40	.20				
	.07	.05	.02	.01				
.4	.20	.40	.20					
	.04	.02	.01					
.3	.20	20						
	.02	.00						
.2	.20							
	.00							

Note: Dash indicates no feasible solution

(iv) The bounds need not center around the true parameter values. For example suppose that the true values are $h^2 = .80$, $e^2 = .10, r = .177, p = 0$. Then the population correlations are $c^* = .90, c = .40$. Entering the tables with those entries we find $.37 \leq h^2 \leq .80$ and $.10 \leq e^2 \leq .15$. (The explanation is clear: the bounds are designed to "just" catch extreme parameter values. If the true parameter values *are* extreme ($p = 0$), then the bounds will indeed barely catch them.) Consequently there is

TABLE 2—UPPER AND LOWER BOUNDS ON ENVIRONMENTABILITY e^2

MZ Twin Correlation c^*	*DZ* Twin Correlation c							
	.1	.2	.3	.4	5	.6	.7	.8
.9	–	–	.12	.15	.20	.40	.60	.80
	–	–	.10	.10	.20	.27	.38	.56
.8	–	.22	.26	.32	.40	.60	.80	
	–	.20	.20	.20	.40	.51	.68	
.7	–	.35	.41	.49	.60	.80		
	–	.30	.30	.40	.60	.75		
.6	.42	.48	.56	.66	.80			
	.40	.40	.40	.60	.80			
.5	.54	.61	.71	.83				
	.50	.50	.60	.80				
.4	.67	.74	.86					
	.60	.60	.80					
.3	.77	.87						
	.70	.80						
.2	.89							
	.80							

Note: Dash indicates no feasible solution.

no justification for focusing on the midpoints of the intervals as point estimates as Behrman and Taubman (p. 437) do. Nor is there any rationale for arbitrarily restricting the range of r^2 to the interval (.003, .11) to reduce the range of uncertainty about the other parameters as Taubman (1976a, p. 866) does.

III. The Causal Model

Up to this point I have accepted Taubman's estimating equations and simply investigated his bounding procedure. Now I wish to inquire into the causal model which led to those estimating equations: what structure produced his reduced form? The question is particularly pressing because he has restrictions on the reduced-form coefficients: compare (4)-(6) with (1)-(3). The answer cannot be found in Taubman's articles for no coherent causal specification is given there.

With respect to determination of genotypes, the structure is clear, being the simplest variant of the standard model of polygenic transmission from parents to children. (See Cyril Burt and Margaret Howard or D. S. Falconer, chs. 7-10.) Let X_1 and X_2 denote the genotypes of father and mother. Their child's genotype is determined by

$$X = X_o + W \tag{11}$$

where $X_o = (X_1 + X_2)/2$ (the midparent genotype) and W (the specific component, representing Mendelian segregation) are independent. In equilibrium, the variances of X_1, X_2, X are equal, so that

$$\sigma^2_{X_o} = (1 + \mu)\sigma^2_X/2, \qquad \sigma^2_W = (1 - \mu)\sigma^2_X/2$$

where $\mu = \sigma_{X_1X_2}/\sigma^2_X$ is the correlation between the parents' genotypes. For another child of the same parents,

$$X' = X_o + W' \tag{12}$$

with W' independent of X_o and (except in the case of MZ twins) of W. Thus for DZ twins, as for ordinary siblings, $\sigma_{XX'} = \sigma^2_{X_o}$, whence

$$g = r_{XX'} = \sigma_{XX'}/\sigma^2_X = (1 + \mu)/2$$

while for MZ twins, $W = W'$, $X = X'$, and so $\sigma^*_{XX'} = \sigma^2_X$. (We continue to use the * to distinguish MZs.) Taubman's random-mating assumption makes $\mu = 0$ and thus $g = 1/2$.[3]

With respect to determination of environments, the structure is quite unclear, there being no well-established specification in the quantitative genetics literature.[4] Let us consider several possibilities.

MODEL 1: Taubman writes, "Since the DZ brothers' environments are correlated, we can write $N' = \rho N + z'$ where z' is a random variable. We then assume that G is uncorrelated with z' or that one DZ brother's genes are uncorrelated with his sib's specific environment" (1976a, p. 860). Translated into my notation, this says $U' = pU + Z'$ where Z' (specific environment) is uncorrelated with X. Symmetry requires that we include the mirror image equation also: $U = pU' + Z$. To permit gene-environment correlation at the individual level while precluding gene-specific-environment correlation across DZ twins, requires that Z' be uncorrelated with X_o as well as with W. These considerations lead us to the specification that the twins' environments are reciprocally determined by

$$U = pU' + Z, \qquad U' = pU + Z' \tag{13}$$

where Z is uncorrelated with X_o and W', and Z' is uncorrelated with X_o and W. Solving this pair of simultaneous equations

[3] This version of the standard genetic model rules out nonadditive gene effects (dominance). When those are present, only a portion of the genotype (the "breeding value") is transmitted according to (11)-(12), another portion (the "dominance deviation") is not transmitted, but is correlated across siblings. Then the slope of child's genotype on midparent genotype drops below 1 (its value in (11)-(12)), being the ratio of additive genotypic variance to total genotypic variance: see Falconer (ch. 7) or Jencks (pp. 270-71). Taubman's interpretation, "Nonadditive gene effects mean that the average [of two variables] . . . is not half their sum" (1976a, p. 860), is of course incorrect.

[4] The unsettled state of environmental transmission in quantitative genetic models is exemplified by Robert Plomin, John de Fries, and John Loehlin.

yields

$$U = q(Z + pZ'), \qquad U' = q(pZ + Z')$$

where $q = 1/(1 - p^2)$. To ensure that $\sigma_U^2 = e^2$ and $\sigma_{UU'} = pe^2$, we set $\sigma_Z^2 = \sigma_{Z'}^2 = e^2/q$ and $\sigma_{ZZ'} = -pe^2/q$. Then, for individuals,

$$\begin{aligned} \sigma_U^2 &= q^2(\sigma_Z^2 + p^2\sigma_{Z'}^2 + 2p\sigma_{ZZ'}) \\ &= qe^2(1 + p^2 - 2p^2) &&= e^2, \\ \sigma_{XU} &= q\sigma_{ZW} &&= rhe \end{aligned}$$

For DZs,

$$\begin{aligned} \sigma_{UU'} &= q^2(2p\sigma_Z^2 + (1 + p^2)\sigma_{ZZ'}) \\ &= qe^2(2p - p(1 + p^2)) &&= pe^2 \\ \sigma_{XU'} &= pq\sigma_{ZW} &&= prhe \end{aligned}$$

Up to here, Taubman's restrictions hold up. But for MZs, $W' = W$, so that

$$\sigma_{XU'}^* = q(p\sigma_{WZ} + \sigma_{WZ'}) = q(1 + p)\sigma_{ZW} = (1 + p)rhe$$

This makes $\sigma_{XU'}^* \neq \sigma_{XU}$, which is absurd. Indeed, with $W' = W$ it is no longer possible for W to be correlated with Z and uncorrelated with Z'. Thus, this attempt to produce Taubman's estimating equations fails for lack of a consistent specification of environmental determination in MZ families.

MODEL 2: (See N. E. Morton.) Suppose that the twins' environments are determined by

$$(14) \quad U = U_o + V, \qquad U' = U_o + V'$$

where U_o (= common environment) and V, V' (= specific environments) are mutually uncorrelated, and also uncorrelated with X_o, W, W', except that $\sigma_{X_oU_o}$ may be nonzero. The idea here is that the parents provide part of the environment (a part which may be correlated with their genotypes), while another part is random. (Note that V, V' are "specific" in a more natural sense than Z, Z' were; they represent deviations from a common component, rather than from a regression of one twin on another.) From (11)–(12) and (14), with the various zero correlation assumptions, we can calculate

$$\begin{aligned} \sigma_U^2 &= \sigma_{U_o}^2 + \sigma_V^2 \\ \sigma_{UU'} &= \sigma_{UU'}^* = \sigma_{U_o}^2 \\ \sigma_{XU} &= \sigma_{XU'} = \sigma_{XU'}^* = \sigma_{X_oU_o} \end{aligned}$$

Consequently

$$(15) \quad p = \sigma_{UU'}/\sigma_U^2 = \sigma_{UU'}^*/\sigma_U^2 = p^*$$

in accordance with Taubman's restrictions. But

$$s = \sigma_{XU'}/(\sigma_X\sigma_U) = \sigma_{XU'}^*/(\sigma_X\sigma_U) = r$$

contrary to his $s = pr$.

For this model, the reduced form (estimating equations) will consist of (4), (5), and

$$(6') \qquad c = h^2/2 + pe^2 + 2rhe$$

A more general variant of this model would allow the specific environments of MZ twins to be correlated, on the grounds that they share more of their activities than DZ twins do. Then $\sigma_{UU'}^* = \sigma_{U_o}^2 + \sigma_{VV'}^* = p^*e^2$, say. The reduced form will then consist of (4),

$$(5') \qquad c^* = h^2 + p^*e^2 + 2rhe$$

and (6'). With 5 unknown parameters (h^2, e^2, r, p^*, p) there is even less identification than in Taubman's scheme: see R. M. Hogarth, Jensen (1975), and the author (1976). Indeed, once p^* is freed from equality with p, one can arbitrarily set $h^2 = 0$ and $r = 0$, and solve the estimating equations to get $e^2 = 1$, $p^* = c^*$, $p = c$. In this solution heritability is zero, the entire excess of MZ over DZ phenotypic resemblance being accounted for by an assumed excess of MZ over DZ environmental resemblance.[5]

MODEL 3: (See G. Chamberlain.) Suppose again that the twin's environments are determined by (14), but now allow σ_{WV}

[5]It is not surprising that the $p = p^*$ assumption plays a crucial role in the twin method, and that it has been the focus of much discussion: see Torsten Husén; Oscar Kempthorne and Richard Osborne; Richard Lewontin (pp. 396–97); Behrman, Taubman, and Wales; the author (1977a). For some evidence which does support $p = p^*$, see S. Scarr, Adam Mathény, Ronald Wilson, and Anne Dolan, and Loehlin and Robert Nichols, ch. 5.

to be nonzero, along with $\sigma_{X_oU_o}$. Thus the specific as well as the common components of an individual's genotype and environment are correlated. We now have

$$\sigma_U^2 = \sigma_{U_o}^2 + \sigma_V^2,$$
$$\sigma_{UU'} = \sigma_{UU'}^* = \sigma_{U_o}^2$$
$$\sigma_{XU} = \sigma_{XU'}^* = \sigma_{X_oU_o} + \sigma_{WV}, \sigma_{XU'} = \sigma_{X_oU_o}$$

Consequently $p^* = p$ as in (15), while

$$s = \sigma_{XU'}/(\sigma_X\sigma_U) = \sigma_{X_oU_o}/(\sigma_X\sigma_U) = dr$$

where $d = \sigma_{X_oU_o}/(\sigma_{X_oU_o} + \sigma_{WV})$. We can obtain $d = p$ and thus Taubman's $s = pr$, by imposing $\sigma_{WV}/\sigma_{X_oU_o} = (1 - p)/p$, which says that the covariances of the specific and common genotype-environment components stand in the same ratio as the variances of the specific and common environmental components.

But that is entirely *ad hoc*, and so Model 3 would only push back the question of environmental determination by introducing an unanalyzed correlation between the specific components of an individual's genotype and his environment. What process produces that correlation?[6]

MODEL 4: Suppose that the specific component of genotype is in fact a determinant of the specific component of environment:

$$U = U_o + bW + V \tag{16}$$
$$U' = U_o + bW' + V'$$

where b is a coefficient (presumably positive) and V, V' are reinterpreted as purely random components of environment, mutually uncorrelated, and uncorrelated also with X_o, U_o, W, W'. We may calculate:

$$\sigma_U^2 = \sigma_{U_o}^2 + b^2\sigma_W^2 + \sigma_V^2$$
$$\sigma_{UU'}^* = \sigma_{U_o}^2 + b^2\sigma_W^2$$
$$\sigma_{UU'} = \sigma_{U_o}^2$$

and need go no further. Clearly, Taubman's assumption of equal environmental correlations is ruled out, since $\sigma_{UU'}^* > \sigma_{UU'}$ implies $p^* > p$. The very process which makes for a heightened gene-environment correlation across *MZs* also makes for a heightened environmental correlation. The same problem arises if we replace W and W' by X and X' in (16).

These four alternative models may not exhaust the possibilities. It is conceivable that some other specification of environmental determination, in particular one which incorporated a special process for *MZ* families, could generate $p^* = p$ along with $s = pr$. But, lacking that, we should conclude that Taubman's environmental restrictions on the reduced form are untenable. His restriction on genotypic transmission is also questionable, since nonrandom mating and nonadditive gene effects would make g deviate from 1/2.[7] Once those restrictions are relaxed, the indeterminancy of the parameters is restored, and we may have to settle for 1 and 0 as the upper and lower bounds for both heritability and environmentability.

IV. Heritability and Policy

Should this indeterminancy be a cause for concern and a stimulus to further research? The answer would be yes if the effectiveness of socioeconomic policies depended on the extent to which the variance in income (or schooling, or occupation) was accounted for by variance in genetic endowments, in particular if high heritability implied that environmental policies were ineffective.

But the h^2, e^2 division is simply a partitioning of observed variance, analogous to

[6]Furthermore, Model 3 would imply additional inequality restrictions which were overlooked in the bounding procedure. The implied squared correlation between W and V is $r_{wv}^2 = 2(1 - p)r^2$. For this to be a proper correlation requires that $2(1 - p)r^2 \leq 1$, an inequality which is violated, for example, at point A in my numerical example for Case 3 (Taubman's *log* earnings data). Similarly the implied squared correlation between X_o and U_o, namely $2pr^2$, must be less than or equal to 1.

[7]Indeed, when g is treated as an unknown parameter in elaborate multivariate analyses of the same data base, Behrman, Taubman, and Wales (p. 60) estimate it to be .35.

an allocation of the multiple R^2 among the explanatory variables in a conventional regression equation. Surely the regression slopes rather than the contributions to R^2 are relevant to assessing the impact of policy changes. To assess the usefulness of a specific policy manipulation, a specific cost-benefit analysis is required; "proportion of variance accounted for" does not provide a useful short-cut (see Glen G. Cain and Harold W. Watts). While the within-twin-pair regression equations reported in Taubman (1976b) are of considerable interest, the heritability estimates seem quite pointless.

Economists might do well to abandon the enterprise of determining the heritability of socioeconomic achievement measures. As Jencks put it: "Indeed our main conclusion after some years of work on this problem is that mathematical estimates of heritability tell us almost nothing about anything important" (p. 76).[8]

APPENDIX

There are a number of cross-references in the Taubman articles which give a misleading impression of the robustness of the analysis. Taubman, upon introducing the assumptions which I have labelled $g = 1/2$, $p^* = p, s = pr$, writes, "...our results are sensitive to several of these assumptions, which unfortunately are not testable within the context of this model. However, the author and Terence Wales have developed a technique which allows us to test these assumptions once we embed the variance component model in a latent variable model" (1976a, p. 860). Referring to the same assumptions, Taubman writes, "...work undertaken subsequent to that reported here gives hope that many of these assumptions can be tested" (1976b, p. 448), and goes on to write that Behrman, Taubman, and Wales "have recently shown that embedding a twin model in a latent variable framework may allow us to distinguish between various models and to identify the separate contributions of genetics and family environment" (1976b, p. 458). But these promises, or hopes, have not been realized. In their multivariate analysis of the same data base, Behrman, Taubman, and Wales do not test the $p^* = p$ and $s = pr$ assumptions, nor are those assumptions testable. Additional arbitrary assumptions are in fact introduced; see also the author (1977a).

There is also some language which is subject to misinterpretation. Behrman and Taubman conclude that "...this study suggests to us that an economic-political system that is basically free enterprise will be one in which economic inequality will be passed on from one generation to another via genetic endowments and family environment" (p. 440), a remark which hints at strong resemblance between parents and children. But there is very little, if anything, in the twin study which speaks to the issue of intergenerational mobility. In particular a high value for h^2 is hardly evidence of immobility. When $h^2 = 1$, the phenotypic parent-child correlation will equal the genotypic parent-child correlation. With random mating and all gene effects additive, that figure would be 1/2: parent's income would account for 25 percent of the variance in child's income. John Conlisk made this point, and also clarified the relationship between equalizing opportunity and increasing heritability.

[8]Heritability analysis has been extensively developed and employed in plant and animal genetics. There it is used to predict the effectiveness of selective breeding programs under constant environmental conditions, not to set limits on the potential effectiveness of environmental improvements. In that context, provided environmental transmission is absent, h^2 (or rather its additive component, "narrow heritability") turns out to be the slope in the regression of offspring's phenotype on parent's phenotype: see Falconer, ch. 11. For discussions of the misuse of heritability estimates in the great IQ debate, see Lewontin, and Marcus Feldman and Lewontin.

REFERENCES

J. Behrman and P. Taubman, "Intergenerational Transmission of Income and Wealth," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 436–40.

———, ———, **and T. Wales**, "Controlling

for and Measuring the Effects of Genetics and Family Environment in Equations for Schooling and Labor Market Success," in Paul Taubman, ed., *Kinometrics*, Amsterdam 1977, 35-96.

C. Burt and M. Howard, "The Multifactorial Theory of Inheritance and its Application to Intelligence," *British J. Statist. Psychol.*, May 1956, *8*, 95-131.

G. G. Cain and H. W. Watts, "Problems in Making Policy Inferences from the Coleman Report," *Amer. Soc. Rev.*, Apr. 1970, *35*, 228-42.

G. Chamberlain, "Are Brothers as Good as Twins?," in Paul Taubman, ed., *Kinometrics*, Amsterdam 1977, 287-97.

J. Conlisk, "Can Equalization of Opportunity Reduce Social Mobility?," *Amer. Econ. Rev.*, Mar. 1974, *64*, 80-90.

D. S. Falconer, *Introduction to Quantitative Genetics*, New York 1960.

M. W. Feldman and R. C. Lewontin, "The Heritability Hang-up," *Science*, Dec. 19, 1975, *190*, 1163-68.

A. S. Goldberger, "On Jensen's Method for Twins," *Educ. Psychol.*, No. 1, 1976, *12*, 79-82.

———, (1977a) "Twin Methods: A Skeptical View," in Paul Taubman, ed., *Kinometrics*, Amsterdam 1977, 299-324.

———, (1977b) "The Genetic Determination of Income," Soc. Syst. Res. Instit. Workshop paper no. 7707, Univ. Wisconsin, July 1977.

Richard J. Herrnstein, *I.Q. in the Meritocracy*, Boston 1973.

R. M. Hogarth, "Monozygotic and Dizygotic Twins Reared Together: Sensitivity of Heritability Estimates," *British J. Math. Statist. Psychol.*, May 1974, *27*, 1-13.

Torsten Husén, *Psychological Twin Research: A Methodological Study*, Stockholm 1959.

Christopher Jencks, *Inequality*, New York 1972.

Arthur R. Jensen, *Genetics and Education*, New York 1972.

———, *Educability and Group Differences*, New York 1973.

———, "The Meaning of Heritability in the Behavioral Sciences," *Educ. Psychol.*, No. 3, 1975, *11*, 171-83.

O. Kempthorne and R. H. Osborne, "The Interpretation of Twin Data," *Amer. J. Hum. Genetics*, No. 3, 1961, *13*, 320-39.

R. C. Lewontin, "Genetic Aspects of Intelligence," in H. L. Roman, et al., eds., *Annual Rev. Genetics:* Vol. 9, Palo Alto 1975, 387-405.

J. C. Loehlin and R. C. Nichols, *Heredity, Environment, and Personality*, Austin 1976.

A. P. Mathény, R. S. Wilson, and A. B. Dolan, "Relations Between Twins' Similarity of Appearance and Behavioral Similarity: Testing an Assumption," *Behav. Genetics*, July 1976, *6*, 343-51.

N. E. Morton, "Analysis of Family Resemblance. I. Introduction," *Amer. J. Hum. Genetics*, May 1974, *26*, 318-30.

R. Plomin, J. C. de Fries, and J. C. Loehlin, "Genotype-Environment Interaction and Correlation in the Analysis of Human Behavior," *Psychol. Bull.*, Mar. 1977, *84*, 309-22.

S. Scarr, "Environmental Bias in Twin Studies," *Eugenics Quart.*, No. 1, 1968, *15*, 34-40.

P. Taubman, (1976a) "The Determinants of Earnings: Genetics, Family, and Other Environments; A Study of White Male Twins," *Amer. Econ. Rev.*, Dec. 1976, *66*, 858-70.

———, (1976b) "Earnings, Education, Genetics, and Environment," *J. Hum. Resources*, Fall 1976, *11*, 447-61.

What We Learn from Estimating the Genetic Contribution to Inequality in Earnings: Reply

By PAUL TAUBMAN*

In his criticism of several articles I have written alone or with colleagues, Arthur Goldberger concentrates on two issues. The first is the statistical methodology that yields our estimate that about 40 percent of the variance in earnings of white males at age 50 is attributable to differences in genetic endowments. He suggests the estimates rely on improper or overly strong assumptions. The second issue is whether in his words, the whole effort is "misguided" and that our results have no implication for policy.

Clearly if the whole effort is misguided, any attempt to understand the statistical issues or to improve the methodology would be barren; hence, I concentrate initially on the question of what one can and cannot learn from (unbiased) estimates of the contribution of genetic endowments to inequality of earnings. There are some extremely important questions that we can answer and other important questions that we cannot answer with this information. Unfortunately people have tried to answer the unanswerable ones in the heated debate in the IQ literature. I believe Goldberger fears that some economists will mistakenly try to use my results to answer the last set of questions.

It is helpful to conduct the analysis within the context of a human capital model in which parents and their children are assumed to invest optimally. We begin by assuming that a person's earnings depend upon his marginal productivity, which is a function of his skills. Let us assume further as numerous writers including Gary Becker and James Meade have done that a person's skills depend upon his genetic endowments (G) and his environment or investments in human capital (N). For simplicity let us also assume that a person's observed phenotypic earnings (Y) are related linearly to his genotype and environment as shown in equation (1).

$$(1) \qquad Y = aG + bN$$

where the coefficients a and b depend on the units in which unobserved G and N are denominated.

We can, of course, calculate the variance of *ln* earnings from an appropriate sample. We can also use equation (1) to express the expected value of this variance in terms of the unobserved variables G and N as

$$(2) \qquad \sigma_Y^2 = a^2\sigma_G^2 + b^2\sigma_N^2 + 2ab\sigma_{GN}$$

Equations (1) and (2) can be used to discuss what questions we can and cannot try to answer with a sample of twins. Let us turn first to the qustions we cannot answer. The debate that followed publication of Arthur Jensen's famous piece on the heritability of IQ and on black-white differences in average IQ focused on three questions: are blacks on average less well endowed genetically with cognitive skills? Can various changes in the environment such as more or better schooling be expected to change the IQ of blacks or whites? And should the government try to override nature's dictates and institute policy changes to alter an individual's or a group's average IQ?

To try to answer these questions, Jensen and others in part relied on studies of twins in which it is possible to obtain estimates of $a^2\sigma_G^2$ and $b^2\sigma_N^2$. It is shown in the Appendix that such information does not allow us to answer the first two questions which require in part an estimate of b, the coefficient on environment in equation (1). The third question involves a value judgment. It is possible that the results obtained from twin

*Professor of economics, University of Pennsylvania.

studies may affect a person's values but there is nothing in the results which demands that nature's dictates should not be overturned.

While there are many interesting questions in the income distribution whose answer depends upon the size of b, there are others that depend upon the size of $a^2\sigma_G^2$ and $b^2\sigma_N^2$. These questions arise because many economists are interested in why inequality occurs and why and how inequality changes. The variance is an oft used measure of inequality. My variance decomposition indicates the broad source of inequality and provides empirical estimates of the coefficients in the model Becker reprints from his earlier Woytinsky Lecture.

The estimates derived are also related to an important policy issue which is not discussed in my 1976 paper but which I wish to discuss here. I concluded my 1976 piece by observing that "Transfer and other programs can be used to achieve greater equality of outcome whether the source of the inequality is genetic, family or other environment" (p. 869-70). The standard economic analysis of income redistribution schemes is couched in terms of both equity and economic efficiency. However, as Arthur Okun and others have noted, if there are capital or other market imperfections, the initial free market solution will be inefficient and it is possible both to redistribute income and to increase economic efficiency. Okun then concludes that while there is no evidence on the extent to which inequality of outcome arises because of inequality of opportunity, he believes the latter inequality is very important. Hence, he hopes that much income redistribution can take place with no loss and perhaps a gain in efficiency.[1]

Data on twins can be used to estimate the contribution of inequality of opportunity, defined more rigorously below, to inequality of outcome and to the general policy issue discussed by Okun. To do this let us assume that all investments in human capital are made by parents for their children.[2] Following Becker, assume that because of capital market imperfections the interest rates charged on investments in human capital are a function of family income. More generally it is assumed that parents maximize a multiperiod utility function in which each child's income is an argument. Among the constraints in the maximization problem is a production function which indicates how genetic endowments and investment in human capital combine to produce skill or earnings capacity. The demand function for investment in human capital will depend on prices (P), the interest rate (i), family income (Y_F), tastes (T), and genetic endowments (G). (See Becker.)

$$I = F(P, i, Y_F, T, G) \tag{3}$$

Suppose we were to express equation (3) as a linear function and to assume that in equation (1) $N = rI + u$ where r is the return on investment and u is a random error that measures the difference between *ex ante* expectations and *ex post* realizations. Substituting this linear equation into (1) and expressing r times the linear coefficients in (3) as new parameters, we have

$$Y = aG + b(cP + di + eY_F + fT + mG + u) \tag{4}$$

We can collect terms and write

$$Y = (a + m)G + bcP + bdi + beY_F + bfT + v \tag{5}$$

where $v = bu$.

In our models we label $(a + m)^2\sigma_G^2$ as the genetic contribution, $[b(cP + di + eY_F + fT]^2$ as the contribution of family or common environment, and σ_v^2 as the contribution of noncommon environment. It is worth emphasizing that in doing so we are counting m, the effect of G on investment in human capital, as attributable to genetic endowments. In the section on statistical

[1] See his discussion beginning on p. 83.

[2] Since we are using a utility-maximization approach, the same arguments will apply if the children make the decisions but parents provide resources.

methodology I argue that this treatment is appropriate for some but not all questions of interest.

I define inequality of opportunity as arising when decision makers face different sets of prices, interest rates, family income, and (when parents are the decision makers for the child's investments) parental tastes. Other economists such as Becker, Howard Bowen, Okun, and R. H. Tawney define inequality of opportunity similarly.[3] I assume that for childhood investments, twins reared together face the same P, i, and Y_F. Parents may prefer one sib to the other, however, a test performed in Jere Behrman, the author, and Terence Wales does not reject the null hypothesis that T does not vary within a family.[4] However P, i, Y_F and T can vary across families and this cross-family variation is the source of inequality of opportunity.

Using the National Academy of Science-National Research Council (*NAS-NRC*) Twin Sample and a more elaborate model than that in my 1976 paper, I estimate (1978) that inequality of opportunity accounts for less than 20 percent of the inequality of outcome among white males. While eliminating 20 percent of inequality of earnings is not trivial, the remaining inequality in earnings is large. Thus those interested in fostering greater equality in the income distribution will not be able to rely solely on policies to reduce inequality of opportunity and will have to face up to the tradeoff issue. Of course, my particular numerical estimate from one sample may be biased for a variety of reasons, some of which are discussed in my 1978 paper. But I submit that twin data can be used to examine an important policy related issue which has not been studied previously. There still remains the issue of the best statistical methodology to use.

[3]John Brittain broadens the definition of inequality of opportunity to include differences arising from genetic endowments. I dislike his definition, since it does not let us distinguish between redistribution schemes which improve or worsen economic efficiency.

[4]Since the null hypothesis tested is whether $(Y_F + T)$ varies within families, we allow for the possibility that the twins' rather than the parents' income determines investments.

I. Statistical Methodology

In his comment Goldberger examines my statistical methodology and the sensitivity of my 1976 estimates to changes in specification of the model and in the data. Many of the issues he examines arise because the model presented in that piece was underidentified even after making some strong assumptions which I consider reasonable. We can identify the parameters in the model by imbedding equation (1) into a latent variable framework.[5] Within such a framework, we estimate that genetic endowments and common environment account for 45 and 12 percent, respectively, of the variance of the *ln* of 1973 earnings of the white males in our sample. Within the latent variable framework, we can estimate several of the parameters to which I assigned values in my 1976 article. However even in the latent variable framework, we continue to assume that the expected value of the cross-twin correlation in the unobserved environment (defined below) is the same for identical and fraternal twins. Since I show below that it is possible to make the estimate of the contribution of genetic endowments equal to zero by assuming a large enough difference in these cross-twin environmental correlations, I will pay particular attention to this assumption. While many of the other issues raised by Goldberger do not apply to the latent variable model or to the estimate of the contribution of genetic endowments and common environment derived from that framework, I will also consider these issues since other researchers may not be able to employ a latent variable model.

Because we will be concerned with the contribution of N and G to the variance of Y, we can simplify notation by normalizing both a and b in equation (1) to 1. Equation (1) applies to an individual. The same equation—with a and b already normalized to 1—also holds for his brother whom we can denote by a prime:

[5]A latent variable is an unobserved variable that is related to two or more observed variables. For an excellent presentation, see Goldberger (1973).

$$(1') \qquad Y'_i = G'_i + N'_i$$

After ordering the twins by their family number, we can calculate cross-twin covariances, σ'_{YY}, and correlations, $\sigma'_{YY}/\sigma_Y\sigma'_Y$, as we would calculate covariances and correlations of any two variables. We can also express these covariances and correlations in terms of the expected values of the unobserved variables as

$$(2') \qquad \sigma'_{YY} = \sigma'_{GG} + \sigma'_{NN} + 2\sigma'_{GN}$$

Equation (2′) applies to any covariance calculated across individuals. As it stands equation (2′) does not help us because it contains 3 unknowns not in equation (2) and only one new observed statistic. My 1976 piece and the latent variable model obtain estimates of parameters by using data on twins and by making certain assumptions.

As noted earlier, Goldberger questions my assumption that the expected value of the cross-twin correlation is the same for identical and fraternal twins. Since this assumption is also made in the latent variable model, it is important to consider it in some detail. To help clarify this issue let me temporarily add as an additional assumption that $\sigma_{GN} = \sigma'_{GN} = 0$. Then denoting identical twins by an asterisk, we can derive the following equations for individuals, identical and fraternal twins:

$$(6) \qquad \textbf{Individuals: } \sigma_Y^2 = \sigma_G^2 + \sigma_N^2$$

$$(7) \qquad \textbf{Identical twins: } \sigma^*_{YY} = \sigma_G^2 + \sigma'_{NN}$$

$$(8) \qquad \textbf{Fraternal twins: } \sigma'_{YY} = \sigma'_{GG} + \sigma'_{NN}$$

In equation (7) we make use of the fact that $G^* = G$. If there is no assortive mating and if all genetic effects are additive, for fraternal twins $\sigma'_{GG} = 1/2\ \sigma_G^2$.[6] Even with this added restriction, equations (6)–(8) form a three-equation system with four unknowns $(\sigma_G^2, \sigma_N^2, \sigma^*_{NN}, \sigma'_{NN})$. One way to close this system is to assume as in my 1976 piece and as in Behrman, Taubman, and Wales that $\sigma^*_{NN} = \sigma'_{NN}$. Notice that if we set $\sigma'_{GG} = 1/2\ \sigma_G^2$ and subtract (8) from (7) we have

$$(9) \qquad \sigma^*_{YY} - \sigma'_{YY} = 1/2\,\sigma_G^2 + (\sigma^*_{NN} - \sigma'_{NN})$$

By assuming that $\sigma^*_{NN} - \sigma'_{NN}$ is zero, we can estimate σ_G^2 as $2(\sigma^*_{YY} - \sigma'_{YY})$. But it is also true that if it is assumed that $(\sigma^*_{NN} - \sigma'_{NN}) = (\sigma^*_{YY} - \sigma'_{YY})$, then the estimated value of σ_G^2 would be zero. A similar condition holds in the latent variable framework.

Thus the issue is whether it is reasonable to assume that $\sigma^*_{NN} = \sigma'_{NN}$. I believe the answer depends upon the question the researcher wishes to examine. Consider the following two questions: 1) How much would inequality be reduced if the skill production function were the same for all individuals and if all parents had the same utility function, faced the same prices, and could borrow unlimited funds at the same rate? 2) How much would inequality be reduced if all children were treated exactly alike? I believe that the first question is the one asked by most economists who write about equality of opportunity. But even if prices, tastes, and skill production functions were the same, parents would invest differently in each child if, as in equation (3), optimal investment in human capital depends on each child's genetic endowments. Thus it is also possible to ask the second question.

I pointed out earlier that in my framework, any effect of G on investment in human capital is counted as being attributable to genetic endowments. But as shown in equation (5), only the common price, family income, and taste terms enter into the category which I label environment. Thus my results can be used to answer the first question, but can only be used to answer the second question in the unlikely event that investments in human capital are not dependent on genetic endowments.

There can be other questions in economics whose answers depend on the size of σ_G^2/σ_Y^2. Whether or not data on twins can be used to answer these questions will depend on whether it is appropriate to use the reduced form models which attribute to G that part of the environment chosen because of G. A case by case examination will be required.

Now let us turn to those points of Gold-

[6]In our latent variable model we estimate σ'_{GG}/σ_G^2 to be .34.

berger's that are applicable to my 1976 piece but not to the latent variable model. Since the system I used was underidentified, having four parameters (σ_G^2, σ_N^2, σ'_{NN}, and σ_{GN}) and three observed statistics, I obtained my estimates by assigning values that spanned the feasible set of one parameter and estimated the other three parameters conditional on each such assigned value. Further I imposed the constraint that the three estimated parameters had to lie within their own feasible set. Goldberger demonstrates that there is another restriction in the system which I overlooked and which will limit further the feasible set of estimated parameters. This is an interesting and correct piece of analysis.

Goldberger also criticizes me for assuming that the specific environment of one fraternal twin is uncorrelated with his sib's genetic endowments. He considers four different behavioral models, none of which generate this *ad hoc* assumption. With the advantage of hindsight and subsequent research, I would now proceed somewhat differently. As noted before, we label or count that part of environment that is caused by genetic differences to be attributable to genetic differences. Since I would also assume that within each family the prices, interest, income, and tastes that enter the investment function are the same for both sibs, I would now opt for the model that says that specific environment is uncorrelated with one own's and one sib's genotype. This assumption would yield different estimates than the ones I presented. Let me add that while I agree that the use of well-specified models helps refine the estimates, I don't think it is inappropriate to use *ad hoc* assumptions when there are competing models.

Finally Goldberger considers how sensitive my results are to changes in the estimated cross-sib correlations. He shows that a combination of a ten point increase and a ten point decrease in the correlations for identical and fraternal twins, respectively, will cause major jumps in the estimated parameters. Now since in my sample the estimated standard errors are about .02, it is very unlikely that both of the estimated cross-twin correlations will differ from their true value by .1 because of sampling variability. Of course, the twins who responded may not be a random subset of the population. We can never rule out the possibility that a bias of such a magnitude will arise either because of the pattern of nonresponses or because of the design of the sample. It is, of course, true that sample design or nonresponses can cause a sample to be atypical and to yield biased estimates. To minimize the effects of biased sample, it is standard practice to rely on replication in other samples. I welcome such research.

The particular estimates generated by the technique used in my 1976 piece are sensitive to certain assumptions. However rather than determining what are the best set of assumptions to use in that framework, I prefer to use the estimates derived from the latent variable model which resolves many of these difficulties. These are that 45 and 12 percent of the variance of *ln* earnings are attributable to genetics and family environment, respectively. If these results stand up under replication, I would also conclude that the most thoroughgoing equality of opportunity policy would have little impact on equality of outcomes for white males. Regardless of its source, the remaining inequality could be reduced by transfer programs or by subsidizing training programs, at a cost of a loss in economic efficiency.

APPENDIX

In this Appendix I demonstrate why estimates on the relative size of the contribution of genotype and environment to the variance of earnings or any other variable is not informative about either the average genotype or environment of various groups or the size of b, the slope coefficient of N in equation (1) in the text.

Since earnings are observed, we can of course calculate its mean and variance, and can express these two observed statistics in terms of their expected values as:

$$\text{(A1)} \quad \bar{Y} = a\bar{G} + b\bar{N}$$

(A2) $\sigma_Y^2 = a^2\sigma_G^2 + b^2\sigma_N^2 + 2ab\sigma_{GN}$

Following Goldberger, define $h^2 = a^2\sigma_G^2/\sigma_Y^2$ and $n^2 = b^2\sigma_G^2/\sigma_Y^2$. Notice that even if we were to set a and b equal to 1, mean earnings depends on $\bar{G}$ and $\bar{N}$ while the variance of earnings depend on the variance and covariance of G and N. It is not difficult to envisage $\bar{N}$ exceeding $\bar{G}$ but σ_G^2 exceeding σ_N^2. Thus knowledge of h^2 is not informative of $\bar{G}/\bar{Y}$.

The difference in average earnings between any two groups B and W is

$$\text{(A3)} \quad \bar{Y}_W - \bar{Y}_B = a(\bar{G}_W - \bar{G}_B) + b(\bar{N}_W - \bar{N}_B)$$

The observed differences in average earnings can be attributed to infinite combinations of differences in $\bar{G}$ or $\bar{N}$. Since h^2 doesn't tell us anything about $\bar{G}/\bar{Y}$, we cannot learn the reasons for (12) being nonzero from data on twins.

Now let us examine why estimates of h^2 that are "large" do not indicate whether or not changes in environment will alter earnings. To do so it is useful to introduce "reaction" functions. Suppose first that genes come only in two varieties, C and D. Now let us separate people by this type. For people with each type of gene we can calculate earnings for each value of N. The distribution of the phenotype for each gene is the reaction function. For simplicity the reaction functions for the C and D gene, given in Figure 1, are specified to have a constant slope of b as in equation (1) in the text.

Now suppose that we could actually assign values of N to individuals. Let N be distributed uniformly between N_1 and N_2 with equal numbers of C's and D's at each N.[7] Then the variance of N will depend upon both the length of $N_2 - N_1$ and the size of b. Given a uniform distribution along N the variance of G depends on the vertical distance between C and D. Since C and D are assumed everywhere parallel, the relative size of h^2 and n^2 depend on the range of environmental variation. That is,

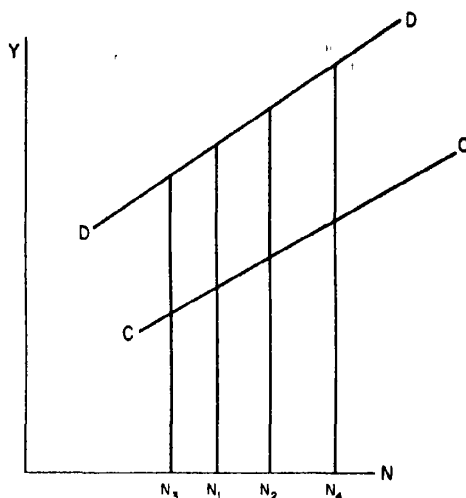



FIGURE 1. TWO HYPOTHETICAL REACTION FUNCTIONS

the relative contribution of N to σ_Y^2 will be greater if the environment is uniformly distributed over $N_4 - N_3$. It may not be as obvious, but Figure 1 also demonstrates than an estimate of σ_G^2 or σ_N^2 is not informative as to whether a change in environment will have large or small effects on earnings. Consider for example a new situation in which the units of N remain unchanged, the slope of the two reaction functions are doubled, but the range of the environment is reduced so that the environmental variance $b^2\sigma_N^2$ remains unchanged. By both rotating the reaction functions and simultaneously narrowing the environmental bounds, we leave both σ_N^2/σ_Y^2 and σ_G^2/σ_Y^2 unchanged. But from equation (1), we know that the effect of a change in N on Y is given by the slope of the reaction function which is greater in the new situation. In practice since we don't observe the reaction functions, we can only estimate the combined term $b^2\sigma_N^2$ and cannot estimate b.

[7] σ_{GN} is thus zero.

REFERENCES

Gary Becker, *Human Capital,* New York 1975.

J. Behrman, P. Taubman, and T. Wales, "Controlling for and Measuring the Effects of

Genetics and Family Environment in Equations for Schooling and Labor Market Success," in Paul Taubman, ed., *Kinometrics: The Determinants of Socioeconomic Success Within and Between Families*, Amsterdam 1977.

Howard Bowen, *Investment in Learning*, San Francisco 1977.

John Brittain, *The Inheritance of Economic Status*, Washington 1977.

A. Goldberger, "The Genetic Determination of Income: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 960–69.

———, "Structural Equation Models: An Overview," in his and Otis Duncan, eds., *Structural Equation Models in the Social Sciences*, New York 1973.

A. Jensen, "How Much Can We Boost IQ and Scholastic Achievement?," *Harvard Educ. Rev.*, No. 1, 1969, *39*, 1-123.

J. Meade, "The Inheritance of Inequalities: Some Biological, Demographic, Social and Economic Factors," *Proc. British Academy*, Dec. 1973, *59*, 1-29.

Arthur Okun, *Equality and Efficiency: The Big Tradeoff*, Washington 1975.

P. Taubman, "The Determinants of Earnings: Genetics, Family and Other Environments; A Study of White Male Twins," *Amer. Econ. Rev.*, Dec. 1976, *66*, 858-70.

———, "Equality of Opportunity and Equality of Outcome, mimeo., Univ. Pennsylvania 1978.

R. H. Tawney, *Equality*, New York 1961.

Income Transfers as a Public Good: Comment

By LAWRENCE SOUTHWICK, JR.*

In an article recently published in this *Review*, Larry Orr develops a model of demand for welfare recipients. He assumes that transfer income to welfare recipients confers a benefit upon taxpayers. From this, he derives the conclusion that the price (the federal matching rate), the incomes of taxpayers, and the number of recipients will be determinants of the level of transfers made.

Following his theory section, Orr estimates a regression equation with the average monthly Aid to Families with Dependent Children (*AFDC*) payment as the dependent variable. He finds that per capita income is positively related to the dependent variable, the proportion of the population receiving aid is negatively related to the dependent variable, the number of recipients is positively related to the dependent variable, there is a negative price effect, and the nonwhite proportion of recipients is negatively related to the dependent variable. Orr then concludes that the empirical estimates are "remarkably consistent with the public good model of income redistribution" (p. 369–70).

While one might well believe in the existence of a demand function for welfare recipients by taxpayers, it is by no means necessary that welfare recipients be a public good. In fact, as will be shown below, a more comprehensive empirical examination of the question tends to cast doubt on the public good characterization.

The major omission by Orr in his estimation of a demand function for welfare recipients lies in the likelihood of the existence of a supply function as well. If the public has a demand for welfare recipients, it is also true that there is a supply of welfare recipients forthcoming.

If the supply of welfare recipients is a function of some of the same variables as the demand, it follows that a simple estimation of the demand function will result in biased estimates for the coefficients. In order to check whether Orr's estimated coefficients are correct, it is necessary to specify the supply function and to estimate the coefficients for both equations in a simultaneous equations framework.

The demand function estimated by Orr takes the average annual *AFDC* payment to a family of four as the dependent variable. His independent variables include per capita income, the recipient/population ratio, the total number of *AFDC* recipients, the marginal nonfederal share of the expense, the fraction of recipient families which are nonwhite, some regional variables, and some other variables which proved less important.

On the supply side, I suggest that the dependent variable is the proportion of the population which is receiving *AFDC* payments. This is likely to be influenced by the benefit level provided and by the unemployment rate. Both should have positive effects; that is, increases in either the unemployment rate or the benefit level should result in a higher proportion of the population being on welfare. The effect of the unemployment rate seems obvious; as it rises, welfare becomes a viable alternative for more people. The effect of the benefit level is twofold. First, as it rises, welfare becomes more attractive. Second, a rise in the benefit level makes more people eligible for welfare benefits.

Other factors which may affect the supply include the income level and the same regional factors used by Orr. The income level again is related to the relative attractiveness of welfare. The regional factors are included for the same reason Orr included them—to see if there are any systematic differences among regions. Finally, the proportion of nonwhites is included simply

*State University of New York-Buffalo. Larry Orr very kindly provided me with his data. I am also indebted to Brian Ratchford for his suggestions. All remaining errors are my responsibility.

TABLE 1—REGRESSION VARIABLES

Variable	Definition
Y	State Per Capita Monthly Income in 1967 Dollars
RPOP	Number of *AFDC* Recipients per 10,000 Population
RCPT	Total *AFDC* Recipients in 1,000's
FED	Marginal Federal Share of *AFDC* Payment, as a Percentage
NWH	Nonwhite *AFDC* Households per 1,000 *AFDC* Households
UNEM	Number of Unemployed per 1,000 Labor Force
BEN	Average Annual Benefit for a Family of Four in 1967 Dollars
NE, OS, BS, W	Dummy Variables for Northeast, Old South, Border States, and West

to see whether it has an effect in either direction.

The definitions of the variables used are given in Table 1. There are some definitions which differ slightly from those used by Orr. Income *Y* is on a monthly basis where Orr uses an annual basis. The reason for the change is to keep all variables in roughly the same order of magnitude. The marginal state share as a fraction has been replaced by the percentage which is the marginal federal share, again for magnitude reasons. Similarly, the fraction of nonwhite households and the recipient/taxpayer ratio have been altered for magnitude reasons.

Of more importance is the fact that the variables of total recipients and the recipient/population ratio are no longer lagged by one year. Orr explains his use of lagged variables as being "in recognition of the likelihood that it takes states some time to adjust to changes in these variables" (p. 365). Of course, if lagged variables are appropriate, the simultaneous equations problem would vanish (to be replaced by a cobweb model). However, policies set by states should be presumed to be in anticipation of the likely consequences (a standard legal presumption). Thus, the state actions are simultaneous with the recipient actions.

Further, the conditions within individual states, both of recipient/population ratios and of benefit levels, are persistent over time. This tends to confirm the hypothesis that the states know what they are doing and do it with intent. It also confirms the simultaneity of the supply and demand functions.

The demand relationship estimated by Orr is given by the equation:

$$(1)\quad BEN = a_0 + a_1 Y + a_2 RPOP + a_3 RCPT + a_4 FED + a_5 NWH + a_6 NE + a_7 W + a_8 OS + a_9 BS + e$$

The supply function which I posit has the form:

$$(2)\quad RPOP = b_0 + b_1 Y + b_2 BEN + b_3 UNEM + b_4 NE + b_5 W + b_6 OS + b_7 BS + b_8 NWH + e$$

Due to the simultaneous equations problem, the reduced form must be estimated first. As pointed out by Orr, *FED* depends on *Y* and *BEN*. Consequently, the general reduced form includes three equations. In these, *BEN*, *RPOP*, and *FED* are determined as functions of all the other variables in the system.

The second-stage regression then uses the first-stage estimated values for these variables where they appear as independent variables in the structural equations (1) and (2). The restated results of Orr's equations (3) and (4) are shown in columns (1) and (2) of Table 2. The results of the two-stage estimation process for the demand function are given in columns (3) and (4) of Table 2. The estimates for the supply function will be presented and discussed later. It should be noted that the two-stage procedure used here is for the purpose of distinguishing between the supply and demand functions

TABLE 2—ESTIMATED REGRESSION COEFFICIENTS AND t-STATISTICS WITH *BEN* AS DEPENDENT VARIABLE-DEMAND FUNCTION

	Columns (3) and (4) from Orr, Table 2[a] 1963–67 (1)	1968–72 (2)	Two-Stage System Estimates 1963–67 (3)	1968–72 (4)
Constant	725.	-9.	2361.	1446.
Y	5.16	7.68	-.95	3.31
	(7.15)	(11.47)	(-.17)	(.46)
RPOP[b]	-.955	-.691	1.860	1.891
	(-4.12)	(3.81)	(1.14)	(.53)
RCPT	.38	.25	.64	-.20
	(2.05)	(2.34)	(.71)	(-.58)
FÊD	4.55	6.72	-33.19	-14.39
	(4.34)	(9.20)	(-.93)	(-.34)
NWH	-.367	-.419	-.497	-1.064
	(-4.00)	(-3.52)	(-4.08)	(-.89)
NE	81.	148.	213.	107.
	(1.57)	(2.32)	(.96)	(1.26)
W	22.	-102.	-181.	-446.
	(.48)	(-1.81)	(-1.82)	(-.84)
OS	-685.	-690.	1079.	-261.
	(8.47)	(-8.22)	(.60)	(-.29)
BS	-137.	-248.	148.	-867.
	(-2.08	(-3.19)	(.23)	(-1.09)
r^2	.73	.78	.68	.70

[a]Orr's constant and coefficients adjusted to be consistent with my variable definitions.

[b]In columns (3) and (4), *RPÔP*.

while the two-stage procedure used by Orr was intended to eliminate simultaneity of demand with the federal matching formula.

In comparing the results for the demand function with those found by Orr, consider first the coefficient for *RCPT*. The positive results Orr found constitute the major justification for calling *AFDC* a public good; the larger the number of recipients, the more benefit an individual taxpayer would receive. From the coefficients in columns (3) and (4), it can be seen that there is no longer a significant positive effect. Thus it cannot be said that *AFDC* is a public good (or a public bad, for that matter, since there is no significant negative effect either).

Orr found a major positive income effect with benefits rising as per capita income rises. This could indicate a desire on the part of the public to share their income gains with the welfare recipients. However, this effect is not evident from columns (3) and (4).

The negative relation between the price (benefit level) and the quantity (recipient/population ratio) was further evidence for Orr of the existence of the demand relationship. However, the coefficients for *RPOP* in columns (3) and (4) are positive, although not significantly so. It might be conjectured that a large ratio of recipients to taxpayers constitutes an effective lobby with the state government. In any case, the demand curve would seem to be quite inelastic.

The effect of the marginal federal subsidy rate is insignificant, although apparently negative, rather than the significant positive effect found by Orr. This is probably because the two-stage procedure used here does not adequately account for the impact of the benefit level on the subsidy rate.

The racial discrimination found by Orr is also evident in the coefficients for *NWH* in columns (3) and (4), although the latter is not significant. It appears probable that

taxpayers are more willing to subsidize white *AFDC* recipients than they are nonwhite recipients.

The regional differences found by Orr appear no longer evident. He found that the Border States (*BS*) and Old South (*OS*) paid lower benefits. However, in columns (3) and (4), these coefficients are no longer significant. It will later be shown that these differences appear in the supply function instead.

Thus far, it appears that my results are largely negative. The results found by Orr are found not to exist in the demand function. However, as we examine the supply function, some interesting and significant results appear. The supply function results are presented in Table 3 where the recipient/population ratio is the dependent variable.

First, note the impact of the benefit level, *BEN*. There is a very strong positive effect in both time periods. In the earlier period, a $1,000 increase in the average annual benefit would cause an increase of 1,887 persons per 10,000 population to be on *AFDC*. In the later period, this effect increased to 2,960 persons. The elasticity in the earlier period was 15.1 and in the later period, 14.9. The effect is probably due to easier eligibility and to migration as well as the greater attractiveness of higher benefits.

Table 3—Estimated Regression Coefficients and *t*-Statistics with *RPOP* as Dependent Variable-Supply Function

	1963-67 (5)	1968-72 (6)
Constant	-1537.	-708.
Y	-7.67	-17.16
	(-8.68)	(-8.05)
BEN	1.887	2.960
	(8.79)	(8.43)
UNEMP	-.58	-3.39
	(-1.15)	(-3.46)
NWH	.86	1.88
	(8.67)	(8.56)
NE	-42.82	-365.77
	(-2.53)	(-6.39)
W	224.48	773.82
	(8.03)	(8.25)
OS	1107.31	1606.07
	(8.87)	(8.86)
BS	753.63	1569.44
	(10.59)	(9.36)
r^2	.48	.49

The effect of income *Y* is strongly negative. The higher the income level, the less attractive a particular level of benefits will be to potential recipients. Further, a higher income level will reduce the number of people eligible to receive benefits. The curve became more income elastic, shifting from -10.2 in 1963-67 to -15.8 in 1968-72.

The effect of the unemployment level is not significant in the earlier period and is negative in 1968-72 with an elasticity of -.45. This is contrary to expectations which indicate that there should be a positive relationship. It may be that this is a spurious effect since most *AFDC* families are single parent and are unlikely to work whatever the unemployment rate.

The proportion of nonwhite households seems to be strongly related to the proportion of the population receiving *AFDC* benefits. The elasticity rose from 1.78 in 1963-67 to 2.45 in 1968-72. A 10 percent increase in the percentage of nonwhite recipient households (averaging 43 percent in the earlier years and 46 percent in the later years) will be accompanied by an 18-25 percent increase in the proportion of the population receiving *AFDC*.

The earlier result of a negative effect of nonwhite households on the benefit level may be a perfectly rational demand response to this supply effect. Since a greater proportion of nonwhite households results in a greater supply of welfare recipients and thus a greater total cost for any particular benefit level, a reduced benefit level may be the only way to keep the total cost within what is felt to be affordable.

In the 1963-67 period, the demand elasticity (benefits with respect to nonwhite households) was -.13. The elasticity of supply (quantity with respect to nonwhite households) was +.18. A 10 percent increase in the nonwhite household proportion would be accompanied by an 18 percent increase in the proportion of the population on *AFDC* and a 13 percent decrease in benefit levels. The latter, of course, would further reduce the supply and

the outcome would be a cost which is in line with what can be afforded. In the 1968–72 period, the comparable elasticities were −.18 and +.25. Again, this suggests that it is cost rather than race prejudice which is at work here.

With regard to the regional differences, a similar explanation may obtain. The supply function shows that the Northeastern states have a lower supply of welfare recipients than other parts of the country while the West has a somewhat higher supply and the Border States and Old South have a substantially higher supply. All of these results are highly significant during both time periods.

It may well be that the Border States and Old South states respond to this higher supply by setting lower benefit levels. Since these states tend to be poorer, this would seem to be logical on the grounds of affordability. These results suggest that the Border States and Old South are simply responding to the greater supply they face along with resultant higher costs by a perfectly rational lowering of benefit levels.

In conclusion, it would appear that many of Orr's conclusions are reduced in significance and others are given reasonable explanations when the interaction of supply with demand is taken into account. In terms of significance, the supply function variables seem to be much better in explaining the observed phenomena than the demand function. Orr has made an interesting first step. This comment represents an enlargement of the system to include a supply function. I do not pretend that it is fully descriptive. Rather, the supply and demand processes for welfare recipients are far more complex than the descriptions given in either this comment or in Orr's paper. Hopefully, an enlarged system of equations can be developed to better describe the interactions within the system and from which policy conclusions can be drawn.

REFERENCES

L. L. Orr, "Income Transfers as a Public Good: An Application to AFDC," *Amer. Econ. Rev.*, June 1976, *66*, 359–71.

Income Transfers as a Public Good: Comment

By BRADLEY R. SCHILLER*

In the June 1976 issue of this *Review*, Larry Orr provided a theoretical framework for ascertaining the optimum level of transfer payments. Building on the foundations of the theory of public goods, he demonstrates that the optimum level of such transfers is a unique function of the (diminishing) marginal utility to taxpayers of improved income support for the poor and the (increasing) marginal cost to taxpayers of larger transfer payments. A unique feature of Orr's approach is his recognition of differences among individuals (taxpayers and recipients alike) in their perceived marginal utilities and costs and his resolution of such differences by simple majority rule. Thus, the optimum level of transfer payments emerges from his model as a unique manifestation of both democratic and economic principles.

Orr utilizes this theoretical framework to rationalize the disparity in welfare benefits available in the different states. The focus of his empirical application is on the Aid to Families with Dependent Children (*AFDC*) program, for the years 1963–67 and 1968–72. In seeking to apply his theoretical framework, Orr must of course identify empirical approximations for his basic arguments. In so doing, I believe he has obscured a critical distinction between racism and classism, and exaggerated the significance of racial prejudices for transfer policy.[1] Although this in no way detracts from the completeness of his model, it is an issue of sufficient concern to merit some further consideration.

I. The Utility Function

As noted above, the optimum level of transfers in Orr's model depends in part on the marginal utility taxpayers derive fron such transfers. It is assumed that the ith taxpayer has a utility function U_i of the form

$$U_i = U_i(Y_i, Y_j)$$

where Y_i is the after-tax income of the ith taxpayer, and Y_j is the vector of after-transfer incomes of recipients. It is further assumed that all recipients receive the same transfer benefit.

The empirical problem, of course, is to identify empirical proxies for the arguments of U_i. In this regard, Orr observes that the level of $\partial U_i/\partial Y_j$ (and therewith optimum benefit levels) will be positively related to both 1) general taxpayer attitudes towards welfare (i.e., "altruistic tastes on the part of taxpayers") and 2) taxpayer attitudes toward specific subpopulations of recipients. Orr provides no empirical counterparts for 1) but does argue that "the racial composition of the caseload [is] ... the most important characteristic upon which taxpayers might discriminate among recipients" (p. 365), and thus that racial composition of the recipient population can be used as an empirical measure of 2). In sum, interstate differences in *AFDC* benefit levels will reflect in part racial prejudices.

II. Interpretation Problems

There are two major problems with Orr's approach. On a pragmatic level, taxpaying whites cannot exercise their racial prejudices without inflicting harm on white recipients, particularly since all recipients in a state are assumed to receive the same benefit. On a more theoretical level, the absence of a measure of general attitudes towards welfare suggests that undue significance may be accorded to racial attitudes. In other words, classism may be disguised as racism in Orr's estimation procedure. In view of the fact that the large coefficients on the race variable lead Orr to conclude that "...rather substantial discrimination

*American University. I wish to thank Larry Orr for his uncommonly generous offer to provide the data and programming from his original study, and Jane Lee for programming.

[1] The distinctive effects of racism and classism on welfare recipients were examined by the author (1971).

against nonwhite recipients on the part of (predominantly white) taxpayers" (p. 368–69), exists, this possibility is worth exploring.

III. General Attitudes Towards Welfare

In order to assess the extent to which Orr's measure of racial phenomena is inflated by general attitudes toward welfare, we need a direct measure of such attitudes. For this purpose I have chosen to use each state's decision on whether or not to make *AFDC* benefits available to families headed by incapacitated males, typically unemployed fathers. This option (*AFDCU*) was created by the 1961 Social Security Act Amendments and elected by the states in subsequent years. Because ability to work is so often regarded as the primary determinant of "deservingness,"[2] the decision to adopt or reject the *AFDCU* program option may be regarded as an empirical expression of a state's general attitude towards welfare. The

[2]See the author (1976, chs. 1, 3, 8, and 12).

TABLE 1—REGRESSION COEFFICIENTS, BY *AFDCU* AND NON-*AFDCU* STRATA

	Orr (1)	1963-67 Non-*AFDCU* States (1a)	*AFDCU* States (1b)	Orr (2)	1968-72 Non-*AFDCU* States (2a)	*AFDCU* States (2b)
Constant	1054.	592.	1880.	677.	-573.	1728.
Y	.42	.22	.62	.63	.65	.69
	(6.95)	(0.12)	(7.01)	(9.59)	(4.44)	(7.33)
RPOP	-8531.	-17220	-8369	-6521.	-3454.	-9964.
	(3.46)	(2.49)	(1.96)	(2.59)	(0.79)	(2.56)
RCPT	.38	3.57	.13	.24	.49	.26
	(2.06)	(2.64)	(0 59)	(2.06)	(0.74)	(1.85)
$\hat{MP}$	-466.	-211.	-347.	-655.	-649.	-756.
	(4.25)	(1.30)	(1.41)	(7.14)	(4.46)	(3.98)
$\hat{F}$	.40	.15	-85.	.09	.46	-.50
	(1.11)	(2.69)	(1.01)	(0.33)	(1.26)	(0.92)
NWH	-308.	229.	-753.	-391.	-242.	-665.
	(2.91)	(0.86)	(4.73)	(2.53)	(0.86)	(2.97)
NE	96.	364.	-23.	143	-123.	266.
	(1.81)	(2.64)	(0.31)	(2.07)	(0.80)	(3.22)
W	18.	114.	-24.	-99.	-158.	-87.
	(0.40)	(0.88)	(0.38)	(1.66)	(1.16)	(1.10)
OS	-568.	-565.	-	-664.	-617.	-
	(4.28)	(3.08)		(5 71)	(3.60)	
BS	-144.	-65	-277.	-249.	-75.	-271.
	(2.18)	(0.41)	(3.07)	(3.06)	(0.46)	(2.44)
T	-148.	-11.	-14.	-27.	143.	-110.
	(2.04)	(0.10)	(0.13)	(0.13)	(0.48)	(0.39)
TSQ	16.	-4.	-4.	1.	-10.	8.
	(1.75)	(0.26)	(0.31)	(0.12)	(0.58)	(0.50)
R^2	.73	.80	.52	.78	.77	.71
N	255.	120.	135.	255.	120.	135.

Notes: *t*-values in parentheses.
Y: State per capita income in 1967 dollars.
RPOP: Ratio of total *AFDC* recipients to state civilian population, lagged one year.
RCPT: Total *AFDC* recipients in 1,000's, lagged one year.
MP: Marginal state share of *AFDC* payments, as a fraction.
F: Federal share of average annual *AFDC* payment to family of four in 1967 dollars.
NWH: Fraction of *AFDC* families headed by nonwhites (multiplied by 1,000).
NE, W, OS, BS: Dummy variables for Northeast, West, Old South, and Border States.
T: Time, in calendar years (1962 = 1).
TSQ: T^2

twenty-eight states that have adopted this option may then be viewed as significantly more "altruistic" than the rest, and therefore likely to have higher $\partial U_t/\partial Y_j$ for any given racial composition. The question addressed here is whether or not racial factors continue to be as significant once this difference in general welfare attitudes is recognized.

To test the sensitivity of Orr's conclusion to this distinction I stratified his sample into *AFDCU* and non-*AFDCU* states and reestimated his regressions.[3] Table 1 summarizes the results of this effort for Orr's original regressions (1) and (2) (other results were comparable). Column (1) depicts Orr's original 1963-67 coefficients for the entire sample of states; columns (1a) and (1b) contain the coefficients estimated for the two subsamples of non-*AFDCU* and *AFDCU* states, respectively; columns (2), (2a), and (2b) are analogous for the period 1968-72.

Because our primary focus here is on the interaction between racial and class attitudes, I will not discuss any coefficients except *NWH*, the fraction of *AFDC* families headed by nonwhites. Nevertheless readers may want to note important variations in other coefficients across strata.[4] With respect to *NWH* itself, note that in both the 1963-67 and 1968-72 periods a higher proportion of blacks in the recipient population appears to have no significant effect on *AFDC* benefit levels in the non-*AFDCU* states. In other words, it appears that the lower benefits available in most non-*AFDCU* states are not so much a function of racial attitudes and proportions as they are a reflection of generally negative perceptions of welfare.[5] In the *AFDCU* states, on the other hand, where general "tastes" for welfare are more altruistic, racial factors do appear to play a significant role. Indeed, racial effects, as measured by the coefficient on *NWH*, are so strong in the *AFDCU* states that they seem to have overwhelmed Orr's initial results, for the composite sample. Thus, it seems appropriate to conclude that although racial attitudes do affect transfer policies, general attitudes toward welfare are important as well, and further, that failure to distinguish between racism and classism may result in distorted policy perceptions.

[3]It may be noted that substantial evidence of heteroscedasticity is apparent in the variances of the two strata, thus lending initial credence to the hypothesis tested here. A dummy variable for *AFDCU* states was also introduced into Orr's equations but did not attain statistical significance.

[4]Notice in particular that no Old South (*OS*) states offer *AFDCU*

[5]Whites comprised 51 2 percent of the *AFDC* caseload in the non-*AFDCU* states in 1962-67 and 49 percent in the later period The simple correlation between *NWH* and *AFDCU* for the entire sample is -.20 in 1963-67 and -.25 in 1968-72; the mean values (and standard deviations) for *NWH* are for 1963-67 .49 (.25) in the non-*AFDCU* states, .38 (.25) in the *AFDCU* states For 1968-72 .51 (27) in the non-*AFDCU* states, and 38 (24) in the *AFDCU* states.

REFERENCES

L. L. Orr, "Income Transfers as a Public Good: An Application to AFDC," *Amer. Econ. Rev.*, June 1976, *66*, 359-71.

Bradley R. Schiller, *The Economics of Poverty and Discrimination*, 2d ed., New York 1976.

———, "Racial Discrimination vs. Class Discrimination," *Rev. Econ. Statist.*, Aug. 1971, *53*, 262-69.

Income Transfers as a Public Good: Comment

By HUGH SPALL*

In a recent article in this *Review*, Larry L. Orr concludes:

> The empirical estimates presented here for the *AFDC* program over the ten-year period 1963–72 are remarkably consistent with the public good model of income redistribution.... The significance of the recipient/population ratio and the absolute number of recipients is especially noteworthy, since these variables are particularly characteristic of the public good model. [p. 369–70]

Orr's empirical results are derived from a three-step estimation procedure. First, the average grant per recipient is estimated using all the variables in the model except the marginal state matching percentage of the *AFDC* grant. Second, the estimated grant from step one is substituted into the matching formula used by the state and the formula is solved for the marginal state percentage. This is the crucial step in the estimating procedure. States are allowed the choice of two federal matching formulas, each of which has different marginal matching percentages. This step of the procedure takes the choice of formula as given, implicitly assuming the choice of formula is not influenced by the size of the grant. The assumption is critical. If the choice of formula is influenced by the size of the grant, the marginal percentage estimated at this step will be subject to simultaneous equation bias and this estimated percentage is carried forward into the third and final step of the estimating procedure. The third step involves estimating the average grant function with the estimated marginal state matching percentage as an explanatory variable.

This comment accomplishes three tasks. First, the size of the grant is shown to affect the choice of matching formula. Second, I demonstrate that treating the choice of formula as exogenous biases the estimated regression coefficients when the public goods model is estimated with 1963–72 data. Third, the public goods model is reestimated with the choice of matching formula incorporated into the estimating procedure.

I

State governments are allowed the choice of two *AFDC* matching formulas. Under the original formula, the state pays one-sixth of the first $18 of the average grant per recipient, a constant percentage of the next $14 of the average grant,[1] and 100 percent of any amount in excess of $32 per recipient. Since 1965, states have been allowed to choose an alternate matching formula. Under this formula, called the Medicaid formula, the state pays a constant percentage of the average grant per recipient. No ceiling exists on federal participation. The implications of both formulas are illustrated by Figure 1.

As Figure 1 indicates, states can reduce their *AFDC* costs by changing matching formulas if the average grant per recipient exceeds a certain critical value. This crucial value (B_c) is defined by:

$$B_c = (-29 + 14P_o)/(P_m - 1) \tag{1}$$

where P_o is the state's percentage of the second $14 of the average grant if the original formula is used and P_m is the marginal state percentage under the Medicaid matching formula. The maximum value of P_m is .50. For most states, B_c exceeds $32 per recipient.

It can now be explained why the use of 1963–72 data will bias the regression coefficient of the marginal state percentage if

*Assistant professor of economics, Central Washington University.

[1] The percentage varies among states.

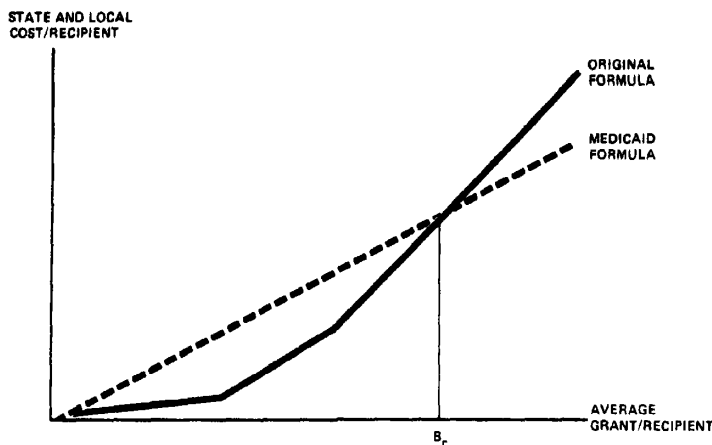



FIGURE 1. THE OPTIMUM *AFDC* MATCHING FORMULA

formula choice is treated as exogenous. Four influences account for the bias. First, states were not allowed to choose between the two matching formulas until 1965. Second, for most states the original formula was more advantageous than the Medicaid formula until the average grant per recipient was $35 or more. Third, between 1965 and 1972, an upward drift in the average grant per recipient caused the average grant to exceed B_c in most states. Fourth, as the average grant per recipient exceeded B_c, states began changing matching formulas and the marginal state percentage declined from 1.0 to the lower Medicaid percentage. As a result, time-series data covering this period will show the average grant per recipient rising while the marginal state percentage is falling. *But note the average grant per recipient is causing the decline in the marginal state percentage, not vice versa.* Even if changes in the marginal state percentage have no causal relationship to average grants, the regression will impute a relationship between the two. Moreover, including time as a variable to control for trend will not prevent this difficulty.

II

If empirical analysis is confined to fiscal year 1975, a one-step extension of the original procedure is sufficient to remove the difficulties presented by the simultaneous relationship between the average grant and the choice of matching formula.[2] The ultimate goal is to estimate

$$(2)\quad B = a_0 + a_1 Y = a_2 RPOP + a_3 RCPT + a_4 MP + a_5 NE + a_6 OS + a_7 BS + a_8 W + a_9 NWH + e$$

where B = the average monthly payment per *AFDC* recipient during fiscal year 1975 (scaled upwards by a factor of 48 for purposes of comparison with the original regression)

[2]If pre-fiscal year 1975 data are used, Orr's estimating technique cannot be corrected by simple one-step addition to the estimating procedure. Prior to January 1974, the same matching percentages had to be used in all categorical assistance programs, not the *AFDC* program alone. The size of the average *AFDC* grant and the number of *AFDC* recipients would have been important determinants because of the relative size of the *AFDC* program, but they would not have been the only determinant. The adoption of the Supplemental Security Income (*SSI*) Program in January 1974 changed this situation. Since the adoption of *SSI*, the other federally funded assistance programs do not affect the choice of *AFDC* matching formula. Thus the simple modification which I suggest will eliminate the bias introduced by the original estimating procedure. If pre-fiscal year 1975 data are used to estimate the model, a one-step extension of the original procedure will not be sufficient to eliminate the bias.

Y = per capita personal income
$RPOP$ = the ratio of $AFDC$ recipients to total state resident population, lagged one fiscal year
$RCPT$ = total $AFDC$ recipients, in thousands, lagged one fiscal year
MP = the marginal state share of $AFDC$ payments (a proportion) during fiscal year 1975
NE, OS, BS, and W are dummy variables representing the Northeast, Old South, Border States, and Western States as defined by the original article.
NWH = the proportion of families who are nonwhite [3]
e = the error term

Equation (2) is identical to regression equation (4) in the original article.

A four-step procedure can be used to estimate equation (2). First, estimate B using B_c and all the variables in equation (2) except MP. B_c is used in estimating B because B_c is a determinant of the matching formula chosen. The elements comprising B_c are exogenous; hence B_c is exogenous. Second, compare the estimated value of B to B_c to see which matching formula is most advantageous to the state. Third, solve the most advantageous formula for MP.[4] Fourth, reestimate equation (2) with the value of MP determined in the previous step serving as an explanatory variable instead of the observed value of MP. This procedure will yield unbiased, although inefficient regression coefficients.

[3]The method used to compute NWH is identical to the method described in the original article. However the data base used is slightly different. Advantage was taken of the 1973 $AFDC$ study which reported NWH for an expanded number of states. Thus fewer states required an estimate of NWH based on regional extrapolations. Where individual state estimates required the use of regional extrapolations, use of the 1973 data permitted the extrapolations to be based on data covering a longer time period.

[4]This step must be eliminated in the case of Arizona since Arizona was not allowed to choose between the two formulas during fiscal year 1975. For Arizona, the original matching formula must be solved for MP.

The results of the estimation are reported in Table 1. For comparison purposes, I also report the results which occur when fiscal year 1975 data are used to estimate the equation with the original procedure. Reporting these results allows us to see how much of the difference is due to different procedures and how much is due to the different data base. Also reported are the empirical results from the original article. Comparing these results with those for fiscal year 1975 allows us to see how well the analytical framework in the original article can explain more recent data.

The numbers without parentheses in Table 1 are regression coefficients. The numbers in parentheses are "t-ratios,"[5] and the numbers in italics are elasticity coefficients evaluated at the means.

An examination of Table 1 raises some interesting points. First, the public goods model of $AFDC$ does not perform as well with fiscal year 1975 data as it does with data from earlier time periods. In particular, the variables most characteristic of the public goods model ($RPOP$ and $RCPT$) have extremely low t-ratios and elasticity coefficients which are uncomfortably close to zero. These results hold for fiscal year 1975 regardless of the estimating technique used. Second, the regional differences captured by the dummy variables appear to have diminished. Except for NE, none of the regional dummies appear to play a significant role. Third, the relationship between B and NWH appears to be less important than in the original study.

Although the results for fiscal year 1975 appear to be at variance with the results obtained from data covering prior years, strong conclusions should not be drawn on the basis of these different results. The original study covered a 10-year period and the results obtained for fiscal year 1975 may reflect an aberration instead of a permanent

[5]The so called t-ratios are not really t distributed. The modified estimating procedure used in this article, as well as the original estimating procedure, constitutes two-stage least squares. The t-ratios are reported for each variable so *relative* significance levels can be compared.

TABLE 1—ESTIMATED REGRESSION COEFFICIENTS

	Equation (4): Original Article	Original Method Applied to Fiscal Year 1975 Data	Modified Method Applied to Fiscal Year 1975 Data
Constant	663	-554	-279
Y	.64	.83	.67
	(11.47)	(5.37)	(3.99)
	1.17	*1.54*	*1.28*
RPOP	-6905	-3828	-2462
	(-3.85)	(-.67)	(-.37)
	-.12	*-.06*	*-.04*
RCPT	.25	.04	.04
	(2.34)	(1.36)	(.96)
	-	*.03*	*.03*
MP	-672	-1288	-155
	(-9.20)	(-3 58)	(-.26)
	-.231	*-.22*	*- 02*
NE	148	564	598
	(2.32)	(2.27)	(2.09)
W	-102	+31	-20.64
	(-1.81)	(+.14)	(-.08)
OS	-690	-376	-559
	(-8.22)	(-1.05)	(-1.38)
BS	-248	-212	-236
	(-3.19)	(-.72)	(-.70)
NWH	-419	-615	-752
	(-3.52)	(-1.09)	(-1.14)
	-	*-.10*	*-.12*
R^2	.78	.74	.67

change in background conditions. Table 1 does suggest the advisability of reestimating the model using data from fiscal year 1976 and 1977 when such data become available to see whether a permanent change in background conditions has occurred. However this is not the main point of this comment.

The main point concerns the degree of bias present in the regression coefficient and *t*-ratio associated with *MP*. Table 1 suggests a significant amount of bias exists. When the method described in the original article is used to estimate equation (2) for fiscal year 1975, *MP* has a "high" *t*-ratio and a respectable elasticity. In fact, the elasticity coefficient is remarkably close to the elasticity coefficient reported in the original article. However, when the revised estimating procedure is used, the significance level of *MP* is dramatically reduced and the elasticity of the variable becomes uncomfortably close to zero. These results are consistent with the changes which should occur when the bias described in Section I is eliminated.

III. Conclusion

On balance, additional evidence appears necessary to substantiate the public good hypothesis. Simultaneous equation bias is present in the original empirical estimates and the revised estimates do not offer strong support for the public good approach. When fiscal year 1975 and the revised estimating procedure are used to estimate the model, the elasticity coefficients of *MP*, *RPOP*, and *RCPT* are close to zero and the *t*-ratios suggest a low degree of statistical significance. Since these variables are characteristic of the public good model, a strong statistically significant relationship should exist between them and *B* before the public good hypothesis is accepted. At present, the main support of the hypothesis is provided by the significance and elasticity coefficient

f *Y*.

A different specification of the model ıight lead to better results. One promising venue of research may be the effect of state nd local tax systems on the level of assis-ınce. Another area might be the effect of ssistance levels in neighboring states. Per-aps if these variables are embodied in the ublic good model, *MP, RPOP,* and *RCPT* ill have the effects postulated by the public ɔod model.

REFERENCES

ın **Kmenta,** *Elements of Econometrics*, New York 1971.

. L. **Orr,** "Income Transfers as a Public Good: An Application to AFDC," *Amer. Econ. Rev.*, June 1976, *66*, 359–71.

U.S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, *Characteristics of State Plans for Aid to Families with Dependent Children, 1974*, Washington 1974.

———, *Findings of the 1967 AFDC Studies*, Washington 1968.

———, *Findings of the 1969 AFCD Studies*, Washington 1970.

———, *Findings of the 1971 AFDC Studies*, Washington 1972.

———, *Findings of the 1973 AFDC Studies*, Washington 1974.

———, *Public Assistance Statistics*, Washington July 1974–June 1975.

U.S. Office of Business Economics, *Surv. Curr. Bus.*, Washington, Apr. 1976.

Income Transfers As A Public Good: Reply

By LARRY L. ORR*

It has been said that the journal article production process is inherently biased toward Type I errors: authors tend to stop running regressions when their *t*-tests reach an "acceptable" level, and journals are loathe to publish insignificant coefficients. If that is true—and I don't doubt it—it is equally true that comments on journal articles are biased toward Type II errors: it is usually easier to produce an insignificant coefficient than it is to produce a better model. A corollary of this observation is, of course, that it is easier to write comments than to write articles.

The authors of the present comments have, through various respecifications of the model presented in my article, produced a whole clutch of insignificant coefficients. The question is, are we dealing here with Type I errors in my model or with Type II errors in those presented in the comments?

In the case of Lawrence Southwick's reestimation of my model with the recipiency rate (*RPOP*) endogenous, the answer is fairly obvious. His explanatory equations for *BEN* (my *B*) aren't identified; each of these equations has two endogenous independent variables (*RPOP* and *FED*) and only one excluded exogenous variable (*UNEMP*). One can therefore place no credence in either the estimated coefficients or their *t*-statistics. His equations (5) and (6), where *RPOP* is the dependent variable, have lots of highly significant coefficients, but these results are simply not believable. They imply, for example, that a $1,000 increase in the Aid to Families with Dependent Children (*AFDC*) benefit level would induce nearly 30 percent of the state population to join the welfare rolls. Given that real benefit levels in this period ranged from less than $400 to over $3,700, and no state ever had more than 10 percent of its population on welfare, that's a little hard to take at face value. I do agree that further research should be done to integrate recipiency rates into the model as an endogenous variable, but Southwick clearly has not produced such a model.

The other two comments present more subtle problems that can't be properly evaluated without additional analysis of the data. Bradley Schiller argues that the underlying structure of benefit determination differs between those states that have adopted *AFDCU* and those that have not, because of differences in "general attitudes toward welfare." In particular, he asserts that this difference in "altruism," rather than racial discrimination, accounts for the significant negative coefficients of the race variable (*NWH*) in my original equations. And indeed, on casual inspection, the separate regressions estimated for *AFDCU* and non-*AFDCU* states in the two time periods do appear to support the notion that a number of the coefficients in my model—including the race coefficient—vary greatly between these two sets of states. Several comments on this apparent result are in order, however.

First, Schiller presents only the coefficient estimates for my equations (1) and (2) for the four subsamples, with the assurance that "other results were comparable." However, the corresponding regressions for my preferred equations (3) and (4) show significant negative coefficients for *NWH* in three of the four subsamples; these coefficients (and their *t*-statistics) are shown in Table 1 below.[1] The only differences be-

*Director, Office of Income Security Policy Research, U.S. Department of Health, Education and Welfare. I wish to thank Lisa Skumatz for programming assistance. All opinions and conclusions contained herein are solely my own, and do not represent the official policy of any agency of the federal government.

[1]The data used for *NWH* in the regressions presented here (and those in Schiller's comment) are slightly modified from those used in my original article. My original data for *NWH* included actual

TABLE 1—COEFFICIENTS OF *NWH*, EQUATIONS (3) AND (4)

	1963-67	1968-72
AFDCU states	−706.	−569.
	(4.63)	(3.00)
Non-*AFDCU* states	−213.	−411.
	(.98)	(2.10)

tween these regressions and those presented by Schiller are the deletion of $\hat{F}$, which was insignificant in three of Schiller's four equations and both of my original equations (1) and (2), and, in 1968-72, deletion of T and *TSQ*, which were also uniformly insignificant in that period.

As a more general test of Schiller's basic proposition of structural differences between *AFDCU* and non-*AFDCU* states, I performed Chow tests on separate regressions (without $\hat{F}$) for the two subsamples in each of the two time periods, as well as *t*-tests for differences in individual coefficients between the two subsamples.[2] The resulting *F*-ratios for structural differences between the two sets of states were 1.24 in 1963-67 and 1.48 in 1968-72, as compared with a critical *F*-value of 1.80 for significance at the 5 percent level. Thus, one cannot reject the hypothesis that the benefit determination equation is the same for both subsamples in each of the time periods.

Moreover, in the 1968-72 period none of the *t*-tests for differences in the individual coefficients except that for the Northeastern regional dummy exceeded 1.0. (The *t*-test for the difference in the coefficient of *NWH* was .42.) In the 1963-67 period, only three of the coefficient differences were significant at the 10 percent level: $Y(t = 2.40)$, $RCPT\ (t = 2.01)$, and $NWH\ (t = 1.85)$. Of course, among twenty-four pairs of coefficients, one would expect two or three to be significantly different at the 10 percent level by chance alone.

On the basis of these statistical tests, I conclude that the evidence for structural differences between the *AFDCU* and non-*AFDCU* states is extremely weak. There is no statistical reason to reject pooling the two subsamples, and interaction of *NWH* with the *AFDCU* dummy is only (weakly) justified in the earlier period. The coefficient instability displayed by Schiller's regressions appears to me to be primarily due to the inclusion of $\hat{F}$ and the marked reduction in sample size involved in splitting the sample.

A couple of final comments on Schiller's results relate to his interpretation of the results. If his argument that "classism" rather than racism explains benefit differentials were valid, I would expect *NWH* to be insignificant in *both* subsamples, not just among non-*AFDCU* states. Taken at face value, his results suggest that racial discrimination is only operative in the "more altruistic" states that have adopted *AFDCU*—a result that I find rather strange.

The data also fail to support his assertions that "The twenty-eight states that have adopted ... [*AFDCU*] may then be viewed as significantly more 'altruistic' than the rest..." (p. 984), and that "... the lower benefits available in most non-*AFDCU* states are ... a reflection of generally negative perceptions of welfare" (p. 984). The most appropriate test of these assertions would appear to be insertion of a simple dummy variable for the presence of *AFDCU*; that is, holding all other variables constant, do mean benefits differ between the two sets of states?[3] As Schiller

values from the 1961, 1967, 1969, and 1971 *AFDC* surveys only for those states for which published values were available, with imputed values for all other states and interpolations or extrapolations for nonsurvey years (see fn 9 of my original paper, p. 365). In the course of further work with these data, I have computed actual values for all states directly from the survey tapes, including the 1973 survey. This should reduce the measurement error in *NWH*, although the sampling variability in some states is rather large. All other variables are identical to those used in my article.

[2]These tests were performed by the method suggested by Ronald Oaxaca.

[3]This intercept-shift interpretation of altruism is consistent with Schiller's remark that the more altruistic states are "therefore likely to have higher $\partial U_i / \partial Y_j$ *for any given racial composition*" (p. 984, emphasis added). It is the same interpretation I placed

himself notes (in his fn. 3), such a dummy was statistically insignificant when added to my original equations. I can report that the *AFDCU* dummy was also insignificant in regressions for each time period in which all of the other independent variables were interacted with *AFDCU* to allow differences in the slope coefficients. In fact, the coefficients of the *AFDCU* dummy in these regressions were *negative* (and rather substantial), rather than positive as Schiller's hypothesis would imply. Thus, there is no empirical support for Schiller's contention that the *AFDCU* states have a generally more positive attitude toward welfare benefit levels.[4]

Hugh Spall argues that my estimate of the effect of the marginal federal matching rate on *AFDC* benefit levels is biased by the simultaneous relationship between benefit levels and the matching rate, and suggests an alternative estimating technique to correct for this bias. Applying his technique to 1975 data he finds that, while my procedure yields results quite consistent with my 1963–72 estimates, his yields a matching rate coefficient that is very small and insignificant.

Spall's argument is that as benefits rose, states found it advantageous to switch from the old *AFDC* matching formula to the Medicaid matching rate: "As a result, time-series data covering this period will show the average grant per recipient rising while the marginal state percentage is falling" (p. 986).

This is really just a restatement of the simultaneity problem addressed in my original article; but it ignores several complicating factors. First, contrary to Spall's assertion, the marginal state percentage (*MP*) does *not* fall monotonically as benefits rise over time. It first rises, as states move along the old *AFDC* formula (below B_c), then falls at the point at which the state opts for the Medicaid formula, then remains constant. Second, much of the variation in both *B* and *MP* in this data base is cross-sectional variation. In cross section, under the *AFDC* formula high-benefit states will obviously tend to have high marginal state matching rates. The same tendency will also be present in cross section under the Medicaid matching rate, since high-benefit states tend to be high-income states, and the federal Medicaid matching rate is negatively related (by statute) to state income.

Thus, the sign of the correlation between *B* and *MP* is an empirical question. In fact, the data do not support Spall's assertion of a negative correlation between *B* and *MP* over the 1963–72 period. In 1963–67, the simple correlation between these two variables was +.79; and in 1968–72 it was +.12. It is hard to believe, then, that the argument advanced by Spall explains the highly significant negative matching rate coefficients obtained in my original regressions. I might also note that nothing in Spall's comment would explain why the matching rate coefficient was significantly negative in the 1963–67 data, where only 7 of 255 observations were for states on the Medicaid formula.

A simple test of Spall's hypothesis for the 1968–72 period is to limit the sample to those states still on the *AFDC* formula. This eliminates the formula switching that Spall cites as the basis for a significant negative matching rate coefficient. For this sample (n = 136), the coefficient of my $\widehat{MP}$ variable is −397 (t = 2.11). There are, of course, problems with using this sample to estimate the structural model, but it nevertheless casts considerable doubt on Spall's explanation.

Nor can Spall's argument explain his own results for 1975. The bias he suggests rests entirely on changes in the matching rate

on the coefficients of the regional dummies in my original article.

[4]It is true that *AFDC* benefits are on average lower in the non-*AFDCU* states, but this fact seems to result as much from lower taxpayer incomes in these states as anything. For example, the mean per capita income differential between *AFDCU* and non- *AFDCU* states in 1968–72 explains 51 percent of the mean *AFDC* benefit differential between those two sets of states, on the basis of my estimated coefficient for *Y*. To the extent that the residual differential does reflect "generally negative perceptions of welfare," these perceptions appear to be better captured by the regional dummies of my original model than by the presence of *AFDCU*.

over time; in a single-year cross section like the one he analyzed, the argument is largely irrelevant. Why then does his technique yield such different results in 1975 from mine? In an attempt to answer that question, I have replicated Spall's method with 1975-76 data and compared it with my own in some detail.

Spall's estimation of $\widehat{MP}$ differs from mine in two respects. In the first-stage estimation of B, he adds the "switch point" B_c as an independent variable. And in the second stage, he assumes that states always take the most advantageous matching formula, given the first-stage estimate of B, whereas I used the formula actually in effect.

The addition of B_c to the first stage appears to have almost no effect on the estimates. Its coefficient in the first stage is insignificant and the resulting second-stage estimates of (Spall's) $\widehat{MP}$ with and without B_c in the first stage are identical for all states except one in each of the two years. In either case, the coefficient of $\widehat{MP}$ (using Spall's second stage) in the regression explaining B is insignificant. In contrast, application of my second-stage technique to the 1975-76 data (deflated to 1967 dollars) yields a coefficient for $\widehat{MP}$ of -655 (t = 3.26), very close to my 1968-72 estimate of -672.[5]

The striking difference in results between the two methods must lie, then, in the different treatments of the second stage. A comparison of the second-stage estimates of $\widehat{MP}$ shows that the two methods give identical results for 87 of the 102 observations in 1975-76. In 12 of the 15 cases where the estimates of $\widehat{MP}$ are different, Spall's method imputes the Medicaid matching rate to a state that is actually on the *AFDC* formula. Since there were only 17 observations of states on the *AFDC* formula to begin with, Spall's method virtually eliminates this matching formula from the data set. In particular, it imputes the Medicaid matching rate to *all* of the observations for states on the *AFDC* formula facing zero federal matching at the margin (MP = 1.0), thereby restricting all but one observation on $\widehat{MP}$ to the narrow interval .25-.50. The effect on the variation of $\widehat{MP}$ is quite striking: the variance of Spall's $\widehat{MP}$ is only about one-sixth the variance of mine. Moreover, within that narrow range, Spall's $\widehat{MP}$ is highly correlated with state per capita income (r = .81), since the Medicaid matching rate is statutorily determined by state per capita income. In contrast, my $\widehat{MP}$ is only weakly correlated with income (r = .13). Given that Spall's method removes most of the independent variation from $\widehat{MP}$ and introduces substantial collinearity with income, it is not surprising that his estimated coefficient of $\widehat{MP}$ is insignificant.

These considerations explain (to me, at least) why Spall obtained an insignificant coefficient for $\widehat{MP}$. They do not, however, settle the question of which approach is correct. We are both agreed that a two-stage approach is necessary, with a first-stage prediction of B in order to eliminate correlation between MP and the error term of the final estimating equation for B. Our difference lies in the second-stage choice of matching rates.

The state's choice of formula is most assuredly endogenous, but the endogeneity is of a peculiar sort. If states were perfectly rational maximizers, they would always be on the "right" formula—and it wouldn't matter whether one used their actual choices or the choices they "ought" to make, as Spall does. Unfortunately, the states are not always on the right formula; typically, states appear to have stayed on the *AFDC* formula for about a year after they passed the "rational" switch point. The question is, does it make sense to impute 60 or 70 percent marginal federal matching to a state that is actually receiving zero matching at the margin? I think not. State *AFDC* benefits undoubtedly take some time to adjust to the federal matching rate; if anything, drastic changes in the matching rate should be lagged in the model, not anticipated.

[5]Spall's estimates for 1975 appear to be in 1975 dollars. Deflating the 1975 matching rate coefficient he obtains with my method to 1967 dollars yields a coefficient of -799.

In summary, then, (a) I find Spall's argument for the existence of bias in my 1963-72 matching rate results quite unconvincing; (b) his argument does not explain either my results for 1963-67 or his own 1975 results, which appear to result merely from the fact that his technique removes most of the independent variation from *MP*; and, (c) on balance, I feel that it makes more sense to treat the choice of matching rate as exogenous than to impute to states marginal matching rates that are radically different from those they actually face. I must admit to being somewhat surprised—and chastened—by the marked sensitivity of the matching rate coefficient to a handful of observations in the 1975-76 period. But that is really just an inescapable result of the almost universal adoption of the Medicaid formula. It seems unlikely that one can learn much about matching rate effects by analyzing states on that formula, because it is so closely related to state income. That does not mean the effects are no longer there—just that our techniques are probably inadequate to disentangle them.

For those readers who have perservered to this point, I present in Table 2 the estimated coefficients for my model based on the 1975-76 data (in 1967 dollars, for comparison with my original estimates).

TABLE 2—ESTIMATED COEFFICIENTS, *AFDC* BENEFIT EQUATION, 1975-76

Independent Variable	Coefficient	t
Constant	292	-
Y	.63	6.95
RPOP	-1197.	.51
RCPT	.30	1.95
MP	-655.	3.26
NWH	-663.	2.76
NE	168.	1.34
W	-187.	1.73
OS	-356.	2.23
BS	-302.	2.12
$\bar{R}^2$	.67	
N	102	

REFERENCE

R. L. **Oaxaca**, "Another Look at Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Amer. Econ.*, Spring 1974, *18*, 23-32.

L. L. **Orr**, "Income Transfers as a Public Good: An Application to AFDC," *Amer. Econ. Rev.*, June 1976, *66*, 359-71.

NOTES

The 1978 Employment Center will be held December 28–30 at the Conrad Hilton Hotel in Chicago, Illinois. Operating hours will be December 28, 10:00 A.M. 5:00 P.M., December 29 and 30, 9:00 A.M.–5:00 P.M. There is no registration fee.

The ninety-second annual meeting of the American Economic Association will be held in Atlanta, Georgia, December 28–30, 1979. The Employment Registry and Center will be open from December 27 to 30.

Call for Papers for the 1979 Meetings

Members wishing to give papers or make suggestions for the program for the meetings to be held in Atlanta, December 28-30, 1979, are invited to send their ideas to Professor Moses Abramovitz, Department of Economics, Stanford University, Stanford, CA 94305. Although most of the sessions sponsored by the American Economic Association will consist of invited papers, there will also be several sessions of noneconometric contributed papers. (The sessions of contributed papers will not be published in the *Papers and Proceedings* issue to appear May 1980.) Proposals for invited sessions should be submitted as soon as possible. To be considered for the contributed sessions, abstracts of proposed (noneconometric) papers must be received no later than March 1, 1979. Economists wishing to give papers on econometrics or economic theory may submit abstracts to the Econometric Society, which meets with the American Economic Association and annually schedules a substantial number of contributions.

Meetings of Regional and International Economic Associations

Eastern Economic Association, last week April 1979, Boston, Massachusetts.

Eastern Finance Association, April 19–21, 1979, Washington, D.C.

History of Economics Society, May 24–26, 1979, University of Illinois, Urbana, IL.

Econometric Society, June 27–30, 1979, Montreal, Quebec, Canada.

Southern Economic Association, November 7–9, 1979, Atlanta, Georgia.

Western Economic Association, June 1979, Las Vegas, Nevada.

Harvard Law School offers four or five Liberal Arts Fellowships to college and university teachers in the arts and sciences for the academic year 1979-80. The purpose of the fellowships is to enable teachers in the social sciences or humanities to study fundamental techniques, concepts, and aims of law, so that, in their teaching and research, they will be better able to use legal materials and legal insights which are relevant to their own disciplines. Fellowship holders will presumably take at least two first-year courses in law, in addition to more advanced courses, and will participate in a joint seminar. The year of study will not count toward a degree. The fellowship grant covers tuition and health fees only. Applications should include a biographical resume (including academic record and list of publications), a statement explaining what the applicant hopes to achieve through his year of study, and two letters of recommendation (mailed to the Chairman directly from the referees). There is no special application form. The deadline for 1979–80 applications is January 15, 1979, submitted to the Chairman, Committee on Liberal Arts Fellowships, Harvard Law School, Cambridge, MA 02138. Awards will be announced before February 15, 1979.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1: for meetings scheduled between November and February, July 1: for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

A conference, "The Trotsky-Stalin Conflict in the 1920s," will be held March 9–10, 1979. For informa-

tion, write UCCIS—University Center for Cultural and Intercultural Studies, Dr. George D. Jackson, Conference Coordinator, Hofstra University, Hempstead, NY 11550.

An International Symposium: Tourism and the Next Decade will be held March 11-15, 1979, at the George Washington University. For information, contact Dr. Donald Hawkins, Steering Committee Coordinator, George Washington University, Tourism Development and Travel Administrative Program, 817 Twenty-third St., NW, Washington, D.C. 20052.

The seventh annual conference of the economics department of the City College of New York, scheduled for May 1979 at the Graduate Center, will be concerned with the current debate over the "New International Economic Order." Keynote addresses will be made by distinguished invited speakers. Proposals for additional papers on any aspect of this subject will be welcomed. Attendance and discussion is open to academic economists, government and United Nations officials, and executives of business firms Inquiries should be addressed to Conference Chairman, Professor Edwin P. Reubens, Department of Economics, City College, City University of New York, Convent Avenue, New York, NY 10031.

The eleventh annual meeting of CHEIRON: The International Society for the History of Behavioral and Social Sciences will be held at the University of Akron, June 8-10, 1979. Papers may deal with aspects of the history of any of the behavioral and social sciences. The emphasis of the meeting will be interdisciplinary. Information concerning the program may be obtained from Dr. Robert G. Weyant, Department of Psychology, University of Calgary, Calgary, Alberta, Canada. Information on local arrangements may be obtained from Dr. John A. Popplestone, Archives of the History of American Psychology, University of Akron, Akron, Ohio 44325.

The U.S. Agency for International Development has established a Studies Division within its newly created Office of Evaluation. The function of the division is to analyze the impact of alternative types of development projects on the income, nutrition, health, and demographic behavior of low income people and the socioeconomic and environmental systems on which their present and future welfare depends. The division hopes to find out what has been learned from past experience by tapping the knowledge of social scientists and development practitioners (especially host country nationals) and the intended beneficiaries of development assistance. While it is expected that the division will conduct field studies, commission reports, and organize workshops, it will not fund basic research.

It would be very helpful to the division's staff if individuals who have had extensive experience with AID projects would write a brief letter giving their name and address, discipline or profession, the countries in which they have observed AID activities in depth, the nature of the activity they have observed, and references to any work they may have published on their experience. Both constructive criticism and observations on projects that seem to be working well are of particular interest. Contact Allan Hoben, Chief, Studies Division, Office of Evaluation, PPC/E, Rm 3673 NS, Agency for International Development, Washington, D.C. 20523.

The Law and Economics Center of the University of Miami School of Law is accepting applications to the John M Olin Fellowship Program in Law and Economics for the class entering September 1979. The three-year program, designed for highly motivated individuals with a strong foundation in graduate microeconomics and its application, leads to the Juris Doctor degree. It prepares Fellows for scholarly research, teaching, legal practice, or government work in the various fields of law and economics. Fellows who have passed the preliminary examinations for the Ph.D. in economics are encouraged to complete the dissertation during their residence in Miami. The program may be completed in two and a half years by attending summer school.

Five fellowships of approximately $33,000 each are available annually. These provide full tuition and fees (currently about $4,000 per year) plus stipends of $7,500 per year for Fellows with the Ph.D. and of $6,500 per year for Fellows without it. Up to five additional fellowships are offered each year. Deadline for submitting applications is February 15, 1979. Awards will be announced by April 1. Candidates must apply separately to the School of Law and should take the Law School Admissions Test not later than December 1978, or arrange to take a special exam no later than January 31, 1979. Individuals interested in more information or in arranging interviews should write the Coordinators, John M. Olin Fellowship Program, Law and Economics Center, University of Miami School of Law, P.O. Box 248000, Coral Gables, FL 33124.

The S.S. Huebner Foundation for Insurance Education is sponsoring its fourth annual research grants competition. The grants are intended to support research in the field of risk and insurance. Full-time faculty members at colleges and universities in the United States and Canada are eligible to apply. Among the areas which will be considered are: risk theory; consumer demand for insurance; risk management; health insurance; other methods of health care

financing; social insurance programs; international insurance issues; insurance law; insurance regulation. Grants are available from the Foundation in amounts up to $10,000. Proposals must be submitted by March 1, 1979, and the grants will be awarded by May 1, 1979. Additional information regarding the program can be obtained from: Dr. J. David Cummins, Research Director, S.S. Huebner Foundation for Insurance Education, W-133 Dietrich Hall/CC, University of Pennsylvania, Philadelphia, PA 19104.

New Journal: Call for Papers

The *Journal of Policy Modeling*, published quarterly by the Society of Policy Modeling, provides a forum for analysis and debates of international policy issues. Focusing upon intertemporal, interregional, and intersectoral policy issues, *JPM* emphasizes formal modeling techniques. Model-oriented, empirically based articles on policy issues are welcome, together with studies about the historical context of current economic, social, and political world problems. Authors and readers are given the opportunity for comments, rebuttals, and rejoinders. Manuscripts should be submitted to: *JPM*, Managing Editor, Box 3299, Grand Central Station, New York, NY 10017.

The American Antiquarian Society will award a number of Visiting Fellowships, June 1, 1979–May 31, 1980, in two categories. *National Endowment for the Humanities Fellowships:* At least two Fellowships with a stipend of $1,666 per month for six to twelve months' residence at the Society, full time. (Fellows may not accept teaching assignments nor undertake any other major fellowships except sabbaticals or grants from their own institutions.) *Fred Harris Daniels Fellowships:* Short-term Fellowships for one to three months, with stipends up to $1,800 maximum. Individuals engaged in scholarly research and writing, including foreign nationals and persons at work on doctoral dissertations may apply.

The deadline for applications for both Fellowships is February 1, 1979. Awards will be announced March 15, 1979. Request application forms from John B. Hence, Research and Publication Officer, American Antiquarian Society, 185 Salisbury Street, Worcester, Massachusetts 01609. Telephone 617+755-5221.

Call for papers: The seventh annual Telecommunications Policy Research Conference (tentatively scheduled for the beginning of May 1979) is now being planned. The conference brings researchers from a variety of disciplines together with policymakers from several branches of government. Those engaged in research which (1) has implications for telecommunications policy, and (2) will be completed by early spring, are invited to submit a *brief* description of their work. If a paper is selected for presentation at the conference, the author will be reimbursed for travel and conference living expenses if no alternative source of funding is available. Please send abstracts *as soon as possible* to: TPRC Organizing Committee; c/o John C. Panzar; Bell Laboratories; Murray Hill, NJ 07974.

Death

Frank Hanna, professor emeritus of economics, Duke University, July 15, 1978.

Retirements

Paul Fisher, chief, International Staff, Social Security Administration, Washington, Aug. 1, 1978.

Werner Hochwald, professor emeritus of political economy, Washington University, St. Louis, July 1, 1978.

Owen H Sauerlender, professor emeritus of economics, Pennsylvania State University, June 30, 1978.

Merton P. Stoltz, provost, Brown University, July 1978.

W. Tate Whitman, Emory University, Sept. 1, 1978.

Visiting Foreign Scholars

Alan Armstrong, University of Bristol, England: visiting professor, department of economics, Duke University, Fall 1978

Richard J. Brook, Massey University, New Zealand: visiting associate professor, Duke University, Jan. 1979.

A. W. Coates, University of Nottingham, England: visiting professor of economics, Emory University, Jan.–June 1979.

Nanak C. Kakwani, University of New South Wales, Australia: visiting professor, department of economics, Wayne State University, Mar.–June 1979.

Patrick C. McMahon, University of Birmingham, England: visiting associate professor, department of economics, Wayne State University, Jan.–June 1979.

Nimrod Mediggo, Tel-Aviv University: visiting associate professor, department of managerial economics and decision sciences, Northwestern University, 1978–79.

Michael G. Porter, Monash University, Australia: visiting professor of economics, Yale University, July 1978.

Fuad Sheikh Salem, University of Jordan: visiting associate professor of management, world business department, American Graduate School of International Management, June 1, 1978.

Maurice F. G. Scott, Oxford University: visiting professor of economics, Yale University, July 1978.

Radha P. Sinha, University of Glasgow, Scotland: visiting professor department of economics, Duke University, Fall 1978.

Daniel Soulie, University of Paris-Dauphine:

visiting associate professor, State University of New York-Stony Brook, Sept. 1978.

Arie Tamir, Tel-Aviv University: visiting associate professor, department of managerial economics and decision sciences, Northwestern University, 1978-79.

Peter J. de la Fosse Wiles, London School of Economics: visiting professor, department of economics, Wayne State University, Mar.-June 1979.

Alan Woodfield, Flinders University of South Australia: visiting professor, department of economics, Duke University, Jan. 1979.

Promotions

Roy D. Adams: associate professor of economics, Iowa State University, Sept. 1, 1978.

Michael J. Applegate: associate professor of economics, Oklahoma State University, Sept. 1, 1978.

Richard Bossert: associate professor of marketing, American Graduate School of International Management, Aug. 23, 1978.

Earl Culp: associate professor of marketing, American Graduate School of International Management, Aug. 23, 1978.

Barry L. Duman: professor of economics, West Texas State University, Spring 1978.

Robert T. Falconer: research officer and senior economist, domestic research department, Federal Reserve Bank of New York, July 20, 1978.

Evelyn C. Fallek: chief, Domestic Banking Application Division, Federal Reserve Bank of New York, Apr. 27, 1978.

Stuart M. Feder: manager, statistics department, Federal Reserve Bank of New York, July 20, 1978.

Richard J. Gelson: manager, statistics department, Federal Reserve Bank of New York, July 20, 1978.

Ken Goldin: professor of economics, California State University-Fullerton, Aug. 28, 1978.

Peter W. Harman: associate professor, department of economics and management, Rhode Island College, July 1, 1978.

Sheng Cheng Hu: professor of economics, Purdue University, Aug. 1978.

Wallace E. Huffman: associate professor of economics, Iowa State University, July 1, 1978.

Ehud Kalai: professor, department of managerial economics and decision sciences, Northwestern University, Sept. 1978.

Marvin Kraus: associate professor of economics, Boston College, Sept. 1978.

Roger M. Kubarych: assistant vice-president, Research and Statistics Function, Federal Reserve Bank of New York, Jan. 1, 1978.

David Levine: associate professor of economics, Yale University, July 1978.

An-loh Lin: associate professor of economics and management, Oakland University, Aug. 1978.

William A. McEachern: associate professor, department of economics, University of Connecticut, Oct. 1978.

Arthur H. Martel: professor of economics, Indiana University of Pennsylvania, Sept. 1978.

Tapan Mitra: associate professor of economics, State University of New York-Stony Brook, Sept. 1978.

Ronald L. Moomaw: associate professor of economics, Oklahoma State University, Sept. 1, 1978.

Wayne Nafziger: professor of economics, Kansas State University, July 1, 1978.

Jon P. Nelson: professor of economics, Pennsylvania State University, July 1, 1978.

Kent W. Olson: associate professor of economics, Oklahoma State University, Sept. 1, 1978.

Donald Ratajczak: professor of economics, Georgia State University, July 1, 1978.

Susan Rose-Ackerman: associate professor of economics and Institution for Social and Policy Studies, Yale University, July 1978.

Mark Rosenzweig: associate professor of economics, Yale University, July 1978.

Mark Satterthwaite: professor, department of managerial economics and decision sciences, Northwestern University, Sept. 1978.

Robert D. Sleeper: manager, foreign department, Federal Reserve Bank of New York, July 20, 1978.

Thomas R. Swartz: professor, department of economics, University of Notre Dame, 1978-79.

Mark Walker: associate professor of economics, State University of New York-Stony Brook, Sept. 1978.

Frederick R. Warren-Boulton: associate professor of economics, Washington University, St. Louis, July 1, 1978.

Dennis J. Weidenaar: professor of economics, Purdue University, Aug. 1978.

Stephen T. Worland: professor, department of economics, University of Notre Dame, 1978-79.

Administrative Appointments

Richard K. Armey: chairman, department of economics, North Texas State University, Sept. 1, 1978.

John O. Blackburn: acting chairman, department of economics, Duke University, 1978-79.

Henry W. Broude: adviser to the president, Yale University, July 1978.

John A. Carlson: chairman, economics policy committee, Purdue University, July 1, 1978.

Michael P. Claudon: chairman, department of economics, Middlebury College, July 1, 1978.

Ward S. Curran: chairman, department of economics, Trinity College, July 1, 1978.

F. Trenery Dolbear, Jr.: chairman, department of economics, Brandeis University, Jan. 1, 1978.

Ken Goldin: acting associate dean, School of Business Administration and Economics, California State University-Fullerton, July 1, 1978.

H. Robert Heller, International Monetary Fund: vice president for international economics, Bank of America, San Francisco, Apr. 1978.

James V. Koch, Illinois State University: dean, Division of Arts and Sciences and professor of economics, Rhode Island College, May 1, 1978.

Richard F. Kosobud: head, department of economics, University of Illinois at Chicago Circle, Sept. 1977.

Gerald M. Lage: head, department of economics, Oklahoma State University, Apr. 1, 1978.

Robert T. Michael: director, National Bureau of Economic Research, Inc., Palo Alto, July 1, 1978.

Merton J. Peck: chairman, economics department, Yale University, July 1978.

Mark B. Schupack: associate dean of the faculty and academic affairs, Brown University, July 1978.

Ashton I. Veramallay: director, Center for Economic Education, Indiana University East, Sept. 1, 1978.

James A. Zwerneman: chairman, department of business administration and economics, Saint Mary's College, Aug. 1978.

Appointments

Kenneth J. Arrow, Harvard University: professor of economics, Stanford University, Sept. 1, 1979.

Ronald M Ayers, Tulane University: lecturer, Ohio State University.

Bruce L Benson, Texas A&M University: visiting assistant professor of economics, Pennsylvania State University, Sept. 1, 1978.

H Woods Bowman: adjunct professor of economics, University of Illinois at Chicago Circle, Spring 1978.

Judith S. Brenneke, Northern Illinois University: assistant professor and director, Center for Economic Education, department of economics and management, Rhode Island College, Sept. 1978

Robert Brusca: economist, Balance of Payments Division, Federal Reserve Bank of New York, Dec 1, 1977.

Roland Buck, Ohio State University: lecturer, department of economics, Ohio State University.

Philip H. Carver, Johns Hopkins University: assistant professor of economics, policy studies program, Dartmouth College, July 1978.

Christophe Chamley, Harvard University: assistant professor of economics, Yale University, July 1978.

Ramaswamy Chandrasekaran, University of Texas-Dallas: visiting professor, department of managerial economics and decision sciences, Northwestern University, 1978–79.

Barry Chiswick: professor of economics, University of Illinois at Chicago Circle, Sept. 1978.

Carmella U. Chiswick: assistant professor of economics, University of Illinois at Chicago Circle, Sept 1978.

Robert E. Christiansen, Indiana University: assistant professor, department of economics, North Texas State University, Sept. 1, 1978.

George Cluff: assistant professor of economics, Georgia State University, Sept. 1978.

David C. Conn, Ohio State University: assistant professor, department of economics, Wayne State University, Sept. 1978.

Robert F. Conrad: assistant professor of economics, Duke University, Fall 1978.

Thomas F. Cosimano, State University of New York-Buffalo: assistant professor of economics, Pennsylvania State University, Sept. 1, 1978.

Charles Craypo, Pennsylvania State University: associate professor, department of economics, University of Notre Dame, Sept. 1978–79.

John T. Cuddington, University of Wisconsin: assistant professor of economics, Stanford University, Sept. 1, 1978.

Charles Dale: financial economist, U.S. Treasury Department, Mar. 29, 1978.

Timothy A. Deyak, Auburn University: visiting assistant professor of economics, University of Iowa, Aug. 29, 1978.

Bruce C. Dieffenback, University of Pennsylvania: associate professor, department of economics, State University of New York-Albany, Sept. 1, 1978.

Stefano Fenoaltea: associate professor, department of economics, Duke University, Fall 1978.

John H. Gates: visiting assistant professor of economics, College of William and Mary, 1978–79.

Amihai Glazer, Yale University: assistant professor of economics, School of Social Sciences, University of California-Irvine, July 1978.

Chennat Gopalakrishnan: professor of social sciences, Law of the Sea Institute, University of Hawaii, May 1978.

Oliver R Grawe: assistant professor of economics, Emory University, Sept. 1978.

Klaus D. Grimm, U.S. Department of Housing and Urban Development: senior fellow, Institute of Manpower Studies, London School of Economics and University of Sussex, England.

Theodore Groves, Northwestern University: professor, department of economics, University of California-San Diego, July 1, 1978.

Luis Guasch, Indiana University: assistant professor, department of economics, University of California-San Diego, July 1, 1978.

Robert E. Hall, Massachusetts Institute of Technology: professor of economics, Stanford University, Sept. 1, 1978.

Richard L. Haney, Jr., University of Georgia: associate professor, Texas A&M University, July 1, 1978.

John Hause, University of Minnesota: professor of economics, State University of New York-Stony Brook, Sept. 1978.

William J. Hausman: visiting assistant professor of economics, College of William and Mary, 1978–79.

James M. Holmes, State University of New York-Buffalo: visiting professor, department of economics, Wayne State University, Sept. 1978–June 1979.

Michael D. Hurd, Stanford University: associate professor of economics, State University of New York-Stony Brook, Sept. 1978.

Curtiss D. Huyser: research associate, department of economics, Iowa State University, Mar. 1, 1978.

Mihailo Ivanovic: economist, Industrial Economies Division, Federal Reserve Bank of New York, Feb. 2, 1978.

John Johnston, University of Manchester: professor of economics, School of Social Sciences, University of California-Irvine, June 1978.

Tahoe Kim: associate professor of economics, American Graduate School of International Manage-

ment, Aug. 23, 1978.

Nancy J. Kimelman, Brown University: instructor, department of economics and management, Rhode Island College, 1978-79.

Susan W. Kramer: visiting assistant professor of economics, College of William and Mary, 1978-79.

Martha Langer: economist, Banking Studies Division, Federal Reserve Bank of New York, Sept. 1978.

Gurcharan S. Laumas, Illinois State University: visiting professor, department of economics, Wayne State University, Sept. 1978-June 1979.

Li Way Lee, Bryn Mawr College: assistant professor, department of economics, Wayne State University, Sept. 1978.

W. Arthur Lewis, Princeton University: visiting professor of social sciences and Afro-American studies, Yale University, July 1978.

Patrick S. McCarthy, Concordia College, Montreal: assistant professor of economics, Purdue University, Fall 1978.

William D. Manson, Virginia Polytechnic Institute and State University: assistant professor, department of economics, Ohio State University

Paul Milgrom, Stanford University: assistant professor, department of managerial economics and decision sciences, Northwestern University, Jan. 1979.

Olivia S. Mitchell, University of Wisconsin-Madison assistant professor, department of labor economics, New York State School of Industrial and Labor Relations, Cornell University, Aug. 1978.

Gene E. Mumy, Virginia Polytechnic Institute and State University: assistant professor, department of economics, Ohio State University.

Richard Murnane, University of Pennsylvania assistant professor of economics, Yale University, July 1978.

William Novshek, Northwestern University: assistant professor of economics, Stanford University, Sept. 1, 1978.

Nai-Pew Ong, Yale University: lecturer, department of economics, Ohio State University.

Craig W. O'Riley: research associate, department of economics, Iowa State University, Mar. 1, 1978.

Mack Ott, Virginia Polytechnic Institute and State University: associate professor of economics, Pennsylvania State University, Sept. 1, 1978.

Richard A. Palfin, Whitman College: assistant professor of economics, Gladys A. Kelce School of Business and Economics, Kansas State College of Pittsburgh.

Thien Pham: economist, Banking Studies Division, Federal Reserve Bank of New York, July 13, 1978.

Michael J. Pogodzinski, State University of New York-Stony Brook: visiting assistant professor of economics, Purdue University, Fall 1978.

Cecile C. Pincince, Clark University: instructor, department of economics and management, Rhode Island College, 1978-79.

Walker Pollard, Virginia Polytechnic Institute and State University: lecturer, department of economics, Ohio State University.

Philip K. Quarcoo, Baruch College: assistant professor, department of economics and management, Rhode Island College, Sept. 1978.

M. Jane Racster: lecturer, department of economics, Ohio State University.

Barbara Lou Reed: instructor of accounting, American Graduate School of International Management, Aug. 23, 1978.

Fred Ribe: economist, Money and Finance Division, Federal Reserve Bank of New York, Sept. 1978.

Randolph E. Ross: visiting associate professor of marketing, American Graduate School of International Management, Aug. 24, 1978.

Alan M. Rugman, University of Winnipeg, visiting associate professor, Graduate School of Business, Columbia University, Aug. 1978-Aug. 1979.

Agnes Rupp: research associate, department of public health, Cornell University Medical College, Mar. 1, 1978.

Jose Sanchez-Molinari: professor, department of economics, College of Saint Teresa, Sept. 1978.

Margaret Schaefer: assistant professor of economics, College of William and Mary, Sept. 1978.

Joseph Shaanan, Cornell University: assistant professor of economics, Oklahoma State University, Sept. 1978.

Gad Shifron, Indiana University. visiting associate professor of economics, Purdue University, Fall 1978.

Frederick T. Sparrow, University of Houston: professor of industrial engineering and economics, Purdue University, Fall 1978.

John W. Speer, U.S. Navy: assistant professor, department of economics and management, Rhode Island College, Sept. 1978.

Daniel F. Spulber, Northwestern University: assistant professor of economics, Brown University, Sept. 1978.

Nancy Stokey, Harvard University: assistant professor, department of managerial economics and decision sciences, Northwestern University, Sept. 1, 1978.

Jerry G. Thursby, Syracuse University: assistant professor, department of economics, Ohio State University.

Marie C. Thursby, Syracuse University: assistant professor, department of economics, Ohio State University.

Charles E. Webster, Jr.: assistant professor of economics, Washington University, St Louis, Sept. 1, 1978.

Arthur L. Welsh, Joint Council on Economic Education: visiting professor of economics, University of Iowa, Aug. 29, 1978.

Robert J. Willis, National Bureau of Economic Research and Stanford University: professor of economics, State University of New York-Stony Brook, Sept. 1978.

Leaves for Special Appointment

Stanley M. Besen, Rice University: codirector and chief economist, network inquiry, Federal Communications Commission, June 1978-June 1980.

Martin Bronfenbrenner, Duke University: visiting fellow, Institute of Development Studies, University of Sussex, Brighton, Fall 1978.

Gary J. Dorman, University of Maryland: U.S. Department of Energy, Policy and Evaluation, Aug. 1978-July 1979.

David I. Fand, Wayne State University: economic advisor to the Comptroller of the Currency, July 1, 1978-June 30, 1979.

David H. Finifter, College of William and Mary: research associate, Brookings Institution and U.S. Department of Labor, Employment and Training Administration, Office of Policy Evaluation and Research, Aug. 1978-July 1979.

Andrew M. Hamer, Georgia State University: consultant, development economics department, World Bank, 1978-79.

Clyde A. Haulman, College of William and Mary: visiting associate professor, department of economics and finance, Florida Technological University, 1978-79

James R Jeffers, University of Iowa: health economics advisor to Agency for International Development, Kenya, and Department of Health, Education, and Welfare, June 1978 Dec. 1979.

Allen C. Kelley, Duke University: visiting research professor, Esmee Fairbairn Research Institute, Heriot-Watt University, Edinburgh, Fall 1978, research scholar, International Institute for Applied Systems Analysis, Laxenburg, Austria, Spring 1979.

Randall R. Kincaid, Jr, Davidson College: senior program analyst, Environmental Protection Agency, June 1978-May 1979.

Steven W. Kohlhagen, University of California-Berkeley. senior staff economist, Council of Economic Advisers, 1978-79.

Charles M. Lucas, Federal Reserve Bank of New York: consultant, International Monetary Fund, Sept. 1978-Aug. 1979.

Bernard L. Weinstein, University of Texas-Dallas: scholar in residence, Southern Growth Policies Board, Washington, D.C., Sept. 1978-Aug. 1979.

E. Roy Weintraub, Duke University: visiting professor, University of Hawaii-Manoa, Fall 1978.

John Weymark, Duke University: visiting assistant professor, University of British Columbia, 1978-79.

Resignations

Robert C. Baesemann, Northwestern University: National Economic Research Associates, Mar. 1978.

Ernst Baltensperger, Ohio State University: University of Heidelberg, June 1978.

Thomas B. Birnberg, Yale University: University of Ottawa, July 1978.

Stephen DeCanio, Yale University: University of California-Santa Barbara, Jan. 1979.

Gary Fields, Yale University: Cornell University, July 1978.

N. Gail Frey, Ohio State University: California State University, Sept. 1978.

Eric Hanushek, Yale University: University of Rochester, July 1978.

T. Krishna Kumar, Florida Atlantic University: Indian Institute of Management, Bangalore, India, Oct. 1, 1978.

Susan Lepper, Yale University: Council of Economics Advisors, July 1978.

Michael L. Lichstein, Ohio State University, Sept. 1978.

Richard Maxon, Iowa State University: Kansas State University, June 7, 1978.

Gerald O'Driscoll, Iowa State University, May 31 1978.

George F. Rhodes, Ohio State University: Colorado State University, Sept. 1978.

Gary Smith, Yale University: University of Houston, July 1978.

Joan Tomlinson, Rhode Island College.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A Please use the following categories.

1- Deaths
2 - Retirements
3—Foreign Scholars (visiting the USA or Canada)
4 - Promotions
5 Administrative Appointments
6—New Appointments
7 - Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9-- Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), and the date at which the change will occur.

C Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

SEVENTY-FIFTH LIST OF DOCTORAL DISSERTATIONS IN POLITICAL ECONOMY IN AMERICAN UNIVERSITIES AND COLLEGES

The present list specifies doctoral degrees conferred during the academic year terminating June 1978. Abstracts will no longer be printed, as they are published by University Microfilms, Ann Arbor, Michigan.

General Economics; including Economic Theory, History of Thought, Methodology, Economic History, and Economic Systems

EILON AMIT, Ph.D. Minnesota 1977. The rate convergence of the core in an economy with production

BARRY L. ANDERSON, Ph.D. Duke 1978. Royal commissions, economists, and policy. A study of economic advisory process in postwar Canada.

ALAN BAQUET, Ph.D. Michigan State 1978 A partial dynamic theory of production, investment, and disinvestment.

CHRISTOPHER BAUM, Ph.D. Michigan 1977 Applications of optimal control theory to macro-economic stabilization policy.

J. HOWARD BEALES, Ph.D. Chicago 1978. The distribution of advertising within an industry.

ROBERT A. BECKER, Ph.D Rochester 1978. Simple dynamic models of equilibrium

PETER BERCK, Ph.D. Massachusetts Institute of Technology 1977. Natural resources in a competitive economy.

ROBERT C. BEVER, Ph.D. Purdue 1978 Process models and production theory. A nonstatistical approach

DAVID BIGMAN, Ph.D. Johns Hopkins 1978. Essays on the neoclassical theory of production and distribution and the theory of aggregation

DAN E. BIRCH, Ph.D. Purdue 1977. Stocks, flows and an integrated model of macro disequilibrium.

TIMOTHY BRENNAN, Ph.D. Wisconsin (Madison) 1978. On the micro foundations of Keynesian macroeconomics.

ELBA K. F. BROWN, Ph.D. Duke 1978. The Keynesian and neoclassical research program: A comparison.

YANG BOO CHOE, Ph.D. Missouri (Columbia) 1977. An essay on the idea and logic of agricultural economics.

UN CHAN CHUNG, Ph.D Princeton 1978. Toward a theory of the price-setting banking firm.

MATTHEW C. COHEN, Ph D. Carnegie-Mellon 1978. Decision making in a committee context: A study of the U.S. House of Representatives Education and Labor Committee deliberations of Title I of the Elementary and Secondary Education Act Extension of 1974.

JAVIER CUENCA, Ph.D. Toronto 1978. The trade and commercial policy of Spain, 1765-1826.

LAWRENCE S. DAVIDSON, Ph.D. North Carolina (Chapel Hill) 1977. The Nixon wage and price controls: A theoretical and empirical analysis.

JAMES L. DEATON, Ph.D. North Carolina State 1977. The adoption and diffusion of the combined harvester-thresher in the United States: A study in economic history

DINESH K. DESAI, Ph.D. Pennsylvania 1978. A macro-econometric model of development.

MARTINE D DUCHATELET, Ph.D. Stanford 1977. Learning by doing and imperfect competition.

TRAN HUU DUNG, Ph.D. Syracuse 1978. Toward a theory of economic activities.

STEPHEN T EASTON, Ph.D. Chicago 1978 Aggregate aspects of the poor law, unemployment insurance, and unemployment in Britain from 1855-1910.

LARRY D ENSMINGER, Ph.D California (Los Angeles) 1977. Brand name capital, a denial Institutions of brand problem genesis.

JAMES S FACKLER, Ph.D Indiana 1977 Measuring inflationary expectations: Theory and tests.

NADIA R. FARAH, Ph.D. Clark 1977 Stabilization policies, rational expectations, and the Phillips curve.

LESLIE FARBER, Ph.D. Minnesota 1977. Price lining: A transaction cost-information cost approach.

JOHN L. FINCH, Ph.D Washington 1977. The economic theory of contingent legal fees

JOHN P FITTS, Ph D. Michigan 1978. The identification of corporate growth motivation.

KATHERINE B. FREEMAN, Ph.D Florida State 1978 The significance of psychological motivation in production and welfare analysis.

RICHARD D. GARRETT, Ph.D. New School 1978. Primitive accumulation in the antebellum cotton South.

JOHN W. GRAHAM III, Ph.D. Northwestern 1978. The composition of household assets and saving over the life cycle. Theory and evidence.

EDWARD J. GREEN, Ph.D. Carnegie-Mellon 1978 Noncooperative games and equilibrium market strategies.

ALAN GREENSPAN, Ph.D New York 1977. Papers on economic theory and policy

ROGER M. HARSTAD, Ph.D. Pennsylvania 1977. Elements of an equilibrium theory of taxation.

DAVID G. HAZARD, Ph.D. California (Riverside) 1978. Accident externalities and the comparative negligence liability rule.

JOHN L. HILLEY, Ph.D. Princeton 1978. An examination of the theory of fairness.

STEPHEN HORNER, Ph.D. Michigan 1977. Stochastic models of technology diffusion.

FRANK HOWARD, Ph.D. Illinois 1977. The theory of long-run industry supply curves: From Marshall to the 1940's.

JAMES M. HVIDDING, Ph.D. Maryland 1976. Microeconomic theories of labor market dynamics and the theory of aggregate demand: Towards an integration.

TADASHI INOUE, Ph.D. Minnesota 1978. On income distribution: The welfare implications of the general equilibrium model, and the stochastic processes of income distribution formation.

MICHAEL JERISON, Ph.D. Wisconsin (Madison) 1977. Optimal public enterprise policies in models of monoplistic competition.

CHANDRA KANODIA, Ph.D. Carnegie-Mellon 1978. Essays on the effects of accounting information on corporate decisions and capital market equilibrium.

ADAM KESSLER, Ph.D. New York 1978. The altruism factors, social space, and the structure of transfer policies

HAK-UN KIM, Ph.D. Pittsburgh 1977 A general equilibrium approach to the long-run equilibrium Phillips curve.

KYOO H. KIM, Ph.D. Wisconsin (Madison) 1978. Capital accumulation with overlapping generations

MELVIN A. KRASNEY, Ph.D. California (Berkeley) 1977. Essays on individual and social welfare

ANN KUSSMAUL, Ph.D. Toronto 1978. Servants in husbandry in early-modern England.

JOHN C. LARSON, Ph.D. Minnesota 1977. Investigation of a new complete system of consumer demand functions

CHIN LIM, Ph.D. Queen's 1978. Uncertainty and implicit contracts.

JAMES T LITTLE, Ph.D. Minnesota 1977 The empirical implications of demand theory.

WILLIAM E. LUNT, Ph.D Stanford 1977. Some formal aspects of the Marxian value analysis

ROBERT A. MCGUIRE, Ph.D. Washington 1978. An empirical investigation of farmers' behavior under uncertainty: Income, price, and yield variability for late nineteenth century American agriculture

LARRY T. MCRAE, Ph.D. North Carolina (Chapel Hill) 1978. Probability confidence and decision theory.

STEPHEN MARTIN, Ph.D Massachusetts Institute of Technology 1977. Two essays on Keynesian user cost.

ABEL M. MATEUS, Ph.D. Pennsylvania 1977. Essays on efficient and optimal economic growth.

STEVEN A. MATTHEWS, Ph.D. California Institute of Technology 1978. Directional and static equilibrium in social decision processes.

JAMES R. MEGINNISS, Ph.D. Chicago 1977. Alternatives to the expected utility rule.

RAJNISH MEHRA, Ph.D. Carnegie-Mellon 1978. Essays in financial economics.

ROGER D. MOREFIELD, Ph.D. Duke 1977. Economic planning and the Economic Council of Canada.

BABU NAHATA, Ph.D. Northern Illinois 1977. Theory of vertical control with variable proportion.

KAZUO NISHIMURA, Ph.D. Rochester 1978. On the problems of uniqueness.

HIROYUKI ODAGIRI, Ph.D. Northwestern 1977. A theory of growth in a corporate economy.

MICHAEL D. PACKARD, Ph.D. Iowa State 1977. An analysis of eleemosynary behavior among individuals.

RICHARD A. PALFIN, Ph.D. Hawaii 1977. The pricing of leasehold and fee simple estates on Oahu: Theoretical considerations and empirical analysis.

ROBERT L. PIROG, Ph.D. Columbia 1978. Resource allocation in a public goods economy with nonconvexities in production.

HAK K PYO, Ph.D. Clark 1977. The life cycle consumption and labor supply under conditions of uncertainty. A theoretical and empirical study.

SUZANNE T. QUINLAN, Ph.D. Georgetown 1978. Risk aversion and the optimal insurance policy.

EDWARD L. RANCK, Ph.D. Kentucky 1977. A refinement of the theory of technological externality with emphasis upon distributional considerations.

JAVIER RUIZ-CASTILLO, Ph.D. Northwestern 1978. Residential choice and general equilibrium theory. Existence and properties of a competitive equilibrium in a special case.

WILLIAM F. SAMUELSON, Ph.D. Harvard 1978. Models of competitive bidding.

WILLIAM E. SCHWORM, Ph.D. Washington 1977. User cost and investment theory.

SHRIDAR SHRIMALI, Ph.D. Southern Illinois 1977. J. P. and the communitarian society: An economic critique.

JOSEPH M. SICILIAN, Ph.D. Purdue 1977. Two essays on externalities and incentive compatibility in welfare economics

FRANCIS J. SPRENG, Ph.D. Pittsburgh 1976. The macroeconomics of Sir Ralph Hawtrey: A mirror image of British economic decline.

NANCY L. STOKEY, Ph.D. Harvard 1978, Life cycle decision making and the theory of fiscal policy.

GEORGE H. SWEENEY, Ph.D. Northwestern 1978. A dynamic theory of the firm subject to regulation.

YITZHAK TAL, Ph.D. Johns Hopkins 1978. Public goods and cooperative solutions in large economies.

MOKHTAR B. TAMIN, Ph.D. Stanford 1978. Rice self-sufficiency in West Malaysia: Micro-economic implications.

MASATSUGU TSUJI, Ph.D. Stanford 1977. Essays on non-Walrasian economics.

RAMAKRISHNA VAITHESWARAN, Ph.D. Iowa State 1978. Economic ethics of Henry Sidgwick.

MICHAEL W. WATTS, Ph.D. Louisiana State 1978. Tudor economic thought after the Reformation: A *Genre* of early English merchantilism.

JEANNE L. WENDEL, Ph.D. Southern Methodist 1977. Theory of a two-stage distribution system.

GERALD A. WHITNEY, Ph.D. Tulane 1977. Government supplied goods and private consumption demand.

ROBERT G. WOLF, JR., Ph.D. Pennsylvania 1978. Competitive markets with asymmetric information.

BEN-ZION ZILBERFARB, Ph.D. Pennsylvania 1977. The impact of indexation on economic stability—A case study of the United States.

Economic Growth and Development; including Economic Planning Theory and Policy, Economic Fluctuations and Forecasting

PHILLIP C ABBOTT, Ph.D. Massachusetts Institute of Technology 1977. Developing countries and international grain trade.

NAZEM M. ABDALLA, Ph.D. Connecticut 1978 Absorptive capacity, foreign capital, and the economic development of Egypt, 1960-72.

ISHER J. AHLUWALIA, Ph.D Massachusetts Institute of Technology 1977. A macro-econometric model of the Indian economy during 1951-73.

AUSAF AHMAD, Ph D. Northern Illinois 1977. The sources of growth and productivity in Indian manufacturing An empirical analysis

KADA AKACEM, Ph D. Colorado 1978 Optimal time control in economic stabilization.

SOHEIL AKHAVAN, Ph.D Florida State 1977. Economic planning in Iran, the fifth development. A simulation model with projection for fifth and sixth national development plans

NAIM H. AL-ADHADH, Ph.D California Institute of Technology 1978 Essays in economic and political choice

MICHEL A AMSALEM, D.B.A Harvard 1978 Technology choice in developing countries The impact of differences in factor costs

EDUARDO U ANINAT, Ph.D Harvard 1977 The redistributive effects of public policy instruments in Chile.

STAVROS APERGIS, Ph.D. California (Los Angeles) 1978. Forecasting Greek agricultural production in a planning framework.

FLORANGELA ARANGO, Ph.D. New York 1978. A dynamic model of unemployment

MOHAMMAD ARIF, Ph.D. Boston 1978. Growth, redistribution and basic needs A case study of Pakistan.

CLARENCE BAYNE, Ph.D. McGill 1977. A study of the oil industry of Trinidad and Tobago as it responds to dynamic changes in the world oil economy.

DAVID B. BELZER, Ph.D. Maryland 1978. An integration of prices, wages, and income flows in an input-output model of the United States.

MATTHEW BERMAN, Ph.D. Yale 1977. Short-run price determination.

HANS G. BICKEL, Ph.D. Maryland 1976. Rice yields and factors involved in their variation among Green Revolution countries and other major rice producing areas.

FARIS T. BINGARADI, Ph.D. California (Riverside) 1977. A study of the external debts of the less developing countries.

OLIVIER BLANCHARD, Ph.D. Massachusetts Institute of Technology 1977. Two essays in economic fluctuations.

HERMINIO BLANCO, Ph.D. Chicago 1978. Investment under uncertainty. An empirical study.

RANDALL BROWN, Ph.D. Wisconsin (Madison) 1977. Productivity, returns, and the structure of production 1947-74.

JOSEPHUS J. C. BRUGGINK, Ph.D. Oregon 1978. Planning for employment in the Third World· The relevance of economic theory for a program of action.

BALU L. BUMB, Ph.D. Maryland 1977. Patterns of rural development in India. An econometric analysis of social, political, and economic change in Rajasthan and Maharashtra.

VIET BURGER, Ph D. Cornell 1978. The economic impact of tourism in Nepal. An input-output analysis.

ROBERT V. BURKE, Ph.D. Toronto 1978 The diffusion of new biological and chemical technologies and the position of small farmers in Mexico.

LUIS R. CACERES, Ph D Utah 1978 Economic integration and underdevelopment in Central America.

JOSHUA R CARPENTER, Ph.D Colorado State 1977 A critique of economic development planning since World War II

P LIZARDO DE LAS CASAS MOYA, Ph.D Iowa State 1977 A theoretical and applied approach towards the formulation of alternative agricultural sector policies in support of the Peruvian agricultural planning process

CHRISTOPHE P. CHAMLEY, Ph.D. Harvard 1977 Aggregate capital accumulation, taxation, and the public debt

PETER T. CHANG, Ph.D. Oklahoma State 1977 A macro-econometric forecasting model of Taiwan

NABIL N. CHARTOUNI, D.B A Harvard 1978. Optimal pricing/investment decisions for natural resources production.

ROMIR CHATTERJEE, Ph.D. New School 1978. Farm mechanization and rural transformation: A study of Shahada Taluka, India.

ARMEANE M. CHOKSI, Ph.D. Minnesota 1978. A planning model for the chemical fertilizer industry.

CAROL CONDON, Ph.D. Columbia 1978. Fluctuations in tax collections of individual states, their explanatory variables, and their relationship to national business fluctuations during the 1950-71 period.

LUIZ A. CORREA DO LAGO, Ph.D. Harvard 1978. The transition from slave to free labor in agriculture in the southern and coffee regions of Brazil: A global and theoretical approach and regional case studies.

ROBERT M. COSTRELL, Ph.D. Harvard 1978. Unempolyment, distribution, and capacity utilization on equilibrium growth paths.

JACQUES C. CREMAR, Ph.D. Massachusetts Institute of Technology 1977. Planning with nondecreasing returns to scale.

BAHRAM DADGOSTAR, Ph.D. Iowa State 1977. Consumer demand for food commodities in Thailand.

MANSOOR DAILAMI, Ph.D. Harvard 1978. The measurement and explanation of man-hour and total factor productivity in a developing economy: A study based on the Iranian manufacturing sector.

M. ARIFEEN DANESHYAR, Ph.D. Southern Illinois 1977. Feasibility of a value-added tax in India.

ALBERT DE PRINCE, Ph.D. New York 1977. A short-run, operationally oriented money stock model amenable to forecasting and control simulations.

KRAIYUTH DHIRATAYAKINANT, Ph.D. California (Los Angeles) 1978. A study of tax structure development: The case of Thailand.

JAMES R. DIGGINS, Ph.D. Harvard 1978. A short-term model of Federal Reserve behavior in the 1970's

ALLEN DRAZEN, Ph.D. Massachusetts Institute of Technology 1977. Essays in the theory of inflation.

ZERUBBABEL O. EBANGIT, Ph.D. Texas (Austin) 1977 The Chinese conception of economic development. Its influence on development strategy in Tanzania.

NATHANIEL O. EJIGA, Ph.D. Cornell 1977. Economic analyses of storage, distribution, and consumption of cowpeas in northern Nigeria

MAHMOUD S EL-FAKERY, Ph.D. Colorado 1978 A simulation model for an oil-based economy: The case of the Socialist People's Libyan Arab Jamahiraya.

HOWARD J. ELLIOTT, Ph.D. Princeton 1977 A benefit-cost analysis of smallholder tree crops in the Ivory Coast

JOSEPH A. FABAYO, Ph D. Purdue 1978. An economic analysis of capacity utilization in selected Nigerian manufacturing industries: 1974-75.

JOSE C. FERREIRA, Ph.D. Vanderbilt 1977. An economic analysis of cassava flour and its effect on nutrition: A case study in Ceara, Brazil

STEPHEN FIGLEWSKI, Ph.D. Massachusetts Institute of Technology 1977. Three essays on information in speculative markets.

JEFFERSON FRANK, Ph.D Yale 1977. A disequilibrium model of inflation and employment.

RODNEY A. FREED, Ph.D. Virginia 1977. The effect of wage-price controls upon the stability of economic equilibrium: An empirical test.

ANTONIO GARCIA-FERRER, Ph.D. California (Berkeley) 1977. Rural internal migration, employment growth, and interregional wage differentials in Spain.

MOHAMMAD A. GHETMIRI, Ph.D. Pennsylvania 1977. Higher oil prices and the optimal development of the Iranian economy.

WILLIAM A. GISBON, Ph.D. California (Berkeley) 1977. The theory of unequal exchange: An empirical approach.

JAGADISH C. GURIA, Ph.D. Boston 1978. A planning model for housing development in a new city.

SANDRA C. HADLER, Ph.D. Maryland 1977. The extended linear expenditure system: An application to Korean household data.

ALAM HAMMAD, D.B.A. George Washington 1977. The development of a system dynamics model for the world petroleum tanker industry.

BASSAM E. HARIK, Ph.D. Wayne State 1978. Economic integration in less developed countries: Prospects for six Arab countries.

PETER W HARMAN, Ph.D. North Carolina (Chapel Hill) 1977. A flow of funds model of a less developed country.

JAMES A. HARRIS, Ph.D. Wayne State 1977. Estimating the effects of financial liberalization on the development of Korea.

MICHAEL D. HARSH, Ph.D. Washington 1978. An economic analysis of the Soviet industrial labor market.

NURHAN HELVACIAN, Ph.D City (New York) 1978. Effects of unpredictable prices on the average unemployment rate.

CLAYTON M. HENRY, Ph.D. Rutgers 1978. Economics of adaption of new farm technology: The case of the Guyanese rice industry

MASAYOSHI HIROTA, Ph.D Rochester 1978. On certain problems of growth economy with many capital goods.

JOHN HUNTER, Ph.D. Wisconsin (Madison) 1977. Land, labor, and capital in agricultural development: A Marxist analysis.

AHMAD JABBARI, Ph.D. Washington (St. Louis) 1977. Employment implications of income redistribution in an input-output framework: Case of Iran.

SALEH JALLAD, Ph.D. Notre Dame 1977. The role of banking in the economic development of Jordan.

JESTH JAOVISIDHA, Ph.D. Connecticut 1978. A short-run macro-economic model for a dualistic economy: The case of Thailand

KHWAJA R. JAVAID, Ph.D. Duke 1978. Mathematical programming and systems analysis applications to planning: A case study of population growth and economic development in Pakistan.

JAMES L. JOHNSON, Ph.D. California (San Diego) 1977. Optimal forecasting and the valuation of securities.

VASSILIOS KALAITZIS, Ph.D. Michigan State 1978. An econometric analysis of the feedgrain cattle economy of Greece.

MALAVANA J. KARUNASEKERA, Ph.D. Cornell 1978. Productivity, technological choice and distribution of income in Taiwan's agricultural sector, 1946-75.

ALBERT KEIDEL III, Ph.D. Harvard 1978. South Korean regional farm product and income 1910-75.

YOUNG S. KIM, Ph.D. Michigan State 1977. Factor substitutability, efficiency growth, and relative wage income shares in the Korean agricultural and manufacturing sectors: 1955–74.

SUK-MO KOO, Ph.D. New York 1978. Demand for risky assets and equilibrium rates of return under uncertain inflation: Construction of multiperiod capital assets pricing model and its application to theory of production and investment.

GEORGE A. KRAFT, Ph.D. North Carolina (Chapel Hill) 1977. Aggregate population models and the theory of economic growth: A synthesis.

LAWRENCE D. KROHN, Ph.D. Columbia 1978. Intertemporal monopoly power.

DANIEL T. LEE, Ph.D. Florida 1977. Technology transfer to developing countries with special reference to the economy of the Republic of China.

YOUNG S. LEE, Ph.D. Maryland 1976. An input-output forecasting model of the Japanese economy.

JOAO E. LIMA, Ph.D. Michigan State 1977. Projections of the impacts of alternative technologies on production and resource allocation in southern Brazilian agriculture, 1970–85.

CHESADA LOOHAWENCHIT, Ph.D. Princeton 1978. A dynamic multicrop model of Thai agriculture: With special reference to the rice premium and agricultural diversification.

THOMAS MCDEVITT, Ph.D. Michigan 1977. On the economic determinants of rural-urban migration in southwestern Nigeria: An exploratory study.

MOHAMMAD G. MADJD, Ph.D. Cornell 1978. Policies concerning sugar production in Iran.

JULIET MAK, Ph.D. Wisconsin (Madison) 1978. Alternative approaches to industrialism: A comparative analysis of the Philippines, South Korea, and Taiwan.

INAYAT MANGLA, Ph.D. Michigan State 1978. Determinants and forecasting of money stock in Pakistan.

HUGO A. MARADIEGUE, Ph.D. Iowa State 1977. A comparative social benefit-cost analysis of the twelve principal projects of Peru's public investment program 1968–75.

MICHAEL A. MARRESE, Ph.D. Pennsylvania 1977. Hungarian investment fluctuations: A theoretical and economic study of hierarchical decision making.

KOOROS MASKOOKI, Ph.D. Nebraska (Lincoln) 1977. Impact of labor scarcity on Iran's economic development: A programming model.

ELKE MELDAU, Ph.D. George Washington 1978. Public health expenditures and income distribution: A case study of Columbia.

KNUT A. MORK, Ph.D. Massachusetts Institute of Technology 1977. Aggregate cost, productivity, and prices in the short run.

PETER MURRELL, Ph.D. Pennsylvania 1977. Long-run economic planning and optimal growth.

WILFRED MWANGI, Ph.D. Michigan State 1978. Farm level derived demand for fertilizer in Kenya.

KATSUYUKI NIIRO, Ph.D. Pittsburgh 1977. A discriminant analysis of the interrelationship between socioeconomic and political variables and changes in total factor productivity.

MEE-KAU NYAW, Ph.D. Simon Fraser 1978. Export expansion and industrial growth in Singapore.

CLAUDIO A. PARDO, Ph.D. Washington 1977. The impact of the external sector on the economic development of Chile and selected Latin American countries.

ALI M. PARHIZGARI, Ph.D. Maryland 1976. Mathematical and econometric models of development planning: The case of Iran.

CHONG S. PARK, Ph.D. Pittsburgh 1976. Domestic consequences of export instability: The case of less developed countries.

JAMES T. PEACH, Ph.D. Texas (Austin) 1978. Land ownership, tenancy, and strategies of economic development in Bangladesh.

TECK PEE, Ph.D. Michigan State 1977. Social returns from rubber research in peninsular Malaysia.

CHAIYUT PILUN-OWAD, Ph.D. New York 1978. The impact of Thailand's economic development plans on foreign trade performance (1958–71): A case study.

SAYED-JAVAD POURMOGHIM, Ph.D. Iowa State 1978. Import demand in developing countries including Iran: A theoretical and empirical study.

MICHAEL D. PRATT, Ph.D. Kansas 1977. Optimal replacement policy with secondhand equipment markets.

MUHAMMAD G. QUIBRIA, Ph.D. Princeton 1978. Foreign resources and economic development: A multisector planning model applied to Bangladesh.

MYUNG RHEE, Ph.D. Michigan State 1977. A systematic study of adjustment time in models of economic growth.

DJAVAD SALEHI-ISFAHANI, Ph.D. Harvard 1977. Population growth and agricultural intensification: A model and some empirical evidence from Iran.

BERNARD A. SCHMITT, Ph.D. Florida State 1977. A cross-section analysis integrating nutritional variables with the economic relationships of developing countries.

CHIRANJIB SEN, Ph.D. Stanford 1978. Essays on the transformation of India's agrarian economy.

HUSHANG SHAHIDI, Ph.D. Colorado State 1977. Economic growth and the distribution of income in Iran.

JAMES SMITH, Ph.D. Michigan 1977. Economy and demography in a Mossi village.

ANDREW B. STOECKEL, Ph.D. Duke 1978. A general equilibrium study of mining and agriculture in the Australian economy.

SHERMAN R. SULLIVAN, Ph.D. New York 1977. Implications of production functions: Yugoslav economic growth, 1952-74.

FERNAN ULATE, Ph.D. Boston 1978. Optimal allocations of resources over time: An application to the Costa Rican case.

GILBERT UWUJAREN, Ph.D. Columbia 1977. A macroeconomic model for development and planning of the Nigerian economy.

PAIROJ VONGVIPANOND, Ph.D. Hawaii 1978. Disaggregation of saving and economic growth: An empirical study for six countries.

GREGORY C. WEEKS, Ph.D. Washington State 1977. The dual labor market, the Phillips curve, and the class conflict business cycle. A synthesis.

NANCY A. WIEGERSMA, Ph.D. Maryland 1976. Land tenure and reform in Viet Nam.

DAVID C. WILCOCK, Ph.D. Michigan State 1978. The political economy of grain marketing and storage in the Sahelian states of West Africa.

CHING-MAI WU, Ph.D Harvard 1977. Property tax incidence and housing markets in Taipei City, Taiwan.

DAVID WYLLIE, Ph D. Connecticut 1978. The impact of a new extractive resource on a declined industrial region: Scotland and the North Sea oil.

ABERA ZEGEYE, Ph.D. Indiana 1978. Price competition between centrally planned economies and less developed countries in the export of primary products to West Europe.

Economic Statistics; including Econometric Methods, Economic and Social Accounting

AHMED ABISOUROUR, Ph.D. Connecticut 1978. An econometric model of the Moroccan economy

KAMRAN AFSHAR, Ph.D. Florida State 1977. A monetary estimate of Iran's *GNP*, 1900-75.

PAUL J. BECK, Ph.D. Texas (Austin) 1977. A critical analysis of the regression estimator in audit sampling.

PAUL G. BENSON, Ph.D. Florida 1977. A Bayesian analysis of model specification uncertainty in forecasting and control.

BRYAN W. BROWN, Ph.D. Pennsylvania 1977. Essays in the identification and estimation of simultaneous equation systems.

CHEN-NAN CHIANG, Ph.D. California (San Diego) 1978. An investigation of the relationships between price series.

LOUIS A. DE NINO, Ph.D. Pittsburgh 1976. An analysis of the interrelationships between the size-distribution of income and economic growth in the United States 1947-73: Equity vs. equality.

JOHN T. FEEZELL, Ph.D. Claremont 1978. Some Markov models of occupational mobility.

TERRY W. FIELDS, Ph.D. Virginia 1978. Time-series analysis, econometric model construction, and a re-examination of the Gibson Paradox.

KISHIN L. GIDWANI, Ph.D. New York 1978. A study in an oil refinery information and control system.

SHARDA A. GUPTA, Ph.D. Purdue 1978. Testing the equality between sets of coefficients in two linear regressions when disturbance variances are unequal.

GABRIEL HAWAWINI, Ph.D. New York 1977. On the time behavior of financial parameters: An investigation of the intervaling effect.

MICHAEL HAZILLA, Ph.D. State University of New York (Binghamton) 1978. The use of economic theory in econometric estimation. Inference in nonlinear constrained models.

RALPH L. HUNTZINGER, Ph.D. Carnegie-Mellon 1978. Market analysis with rational expectations: Theory and estimation.

ROBERT G. JAMES, Ph.D. Oregon 1978. The permanent income hypothesis and other aggregate consumption functions.

BOYAN JOVANOVIC, Ph.D. Chicago 1978. Job matching and the theory of turnover.

DAVID C KLEINMAN, Ph.D. Chicago 1977. Statistical inference in stable distributions based on asymptotic normality of the sample characteristic function.

JESSE M. LEVY, Ph.D. North Carolina (Chapel Hill) 1977. A comparison of continuous and grouped logit estimation procedures.

ANTHONY LIBERATORE, Ph.D. Connecticut 1978. An econometric model of the Connecticut economy.

RAPHAEL J. MICHALSKI, Ph.D. Iowa State 1977. An application of consistent statistical estimation of a non-linear macro-economic policy model.

MOHAMMAD J. MOJARRAD, Ph.D. Pennsylvania 1977. The application of comparative Monte Carlo methods to econometrics: An efficient Monte Carlo study of finite sample properties of iterative instrumental variables estimation.

GEORGE E MONAHAN, Ph.D. Northwestern 1977. On optimal stopping in a partially observable Markov process with costly information.

JAMES N. MORGAN, Ph.D. Missouri (Columbia) 1978. A quarterly econometric model of Missouri.

MICHAEL MOHR, Ph.D. Lehigh 1977. A quarterly econometric model of the long-term structure of production, factor demand, and factor productivity in ten *U S* manufacturing industries

HAROLD L. NELSON, Ph.D. California (San Diego) 1977. The use of Box-Cox transformation in economic time-series analysis: An empirical study.

JEROME A. OLSON, Ph.D. North Carolina (Chapel Hill) 1977. Small sample properties of estimators for stochastic frontier production functions.

ARIE OVADIA, Ph.D. Pennsylvania 1978. Inventory

holding and expected inflation—An empirical study of three industries.

LAWRENCE RADECKI, Ph.D. Michigan 1977. Simultaneous equation models with lagged endogenous variables and autoregressive disturbances.

RONALD RONG-SHENG WANG, Ph.D. New York 1977. Estimation in the presence of sequential parameter variation of a mixed-parameter regression model.

WILLIAM B. STANLEY, Ph.D. Florida State 1977. An econometric analysis of personal injury and wrongful death litigation.

JORGE I. VELEZ, Ph.D. Florida 1978. Bayesian modeling of nonstationarity for lognormal processes.

ASAD ZAMAN, Ph.D. Stanford 1978. The single period control problem in econometrics.

Monetary and Fiscal Theory, Policy, and Institutions

PAUL J. ABBONDANTE, Ph.D. Virginia Polytechnic Institute 1978. Variable risk and the term structure

CHRISTOPHER M. ADAM, Ph.D. Harvard 1977 Decision processes in monetary policy.

RONALD G. ALLAN, Ph.D. George Washington 1977. The demand for money in France: An econometric analysis.

GARY R. ALLEN, Ph.D. Virginia 1978. An investigation of the combined effects of property taxes and local public spending on property values· The case of Virginia.

STUART D. ALLEN, Ph.D. Virginia 1977. The causation of inflation in Switzerland, 1952-75.

SEBASTIAN ARANGO, Ph.D. New York 1977. A portfolio approach to the demand for money in an open economy.

JAY M. ATKINSON, Ph.D. Virginia Polytechnic Institute 1978. A study of regulatory goals and controls: Firm size in the savings and loan industry.

ALAN J. AUERBACH, Ph.D. Harvard 1978. Essays on the taxation of capital income.

DALE G. BAILS, Ph.D. Nebraska (Lincoln) 1977. Econometric tax model of Nebraska.

TOMAS BALINO, Ph.D. Chicago 1977. Argentine monetary and banking reform of 1946.

ANDY H. BARNETT, Ph.D. Virginia 1978. Taxation for the control of externalities.

DONNA K. B. BARNHILL, Ph.D. Texas (Austin) 1977. Arguments on petroleum company divestiture: An evaluation.

JAMES L. BEAVER, Ph.D. Virginia 1977. Monetary policy and the money supply of Sweden.

DAVID K. H. BEGG, Ph.D. Massachusetts Institute of Technology 1977. Rational expectations of inflation and the behavior of asset returns in a stochastic macro model.

RAMESH C. BHATIA, Ph.D. West Virginia 1978. Banking structure and performance: Case study of the Indian banking system, 1950-68.

CHARLES T. BRUMFIELD, Jr., Ph.D. South Carolina 1978. James J. Saxon's influence on commercial banking in the United States.

VICTOR CANTO, Ph.D. Chicago 1977. Taxation, welfare, and economic activity.

PHILIP CARUSO, Ph.D. Michigan State 1977. The impact of skewness of the income distribution on local educational expenditures.

ATHANASSIOS F CATSAMBAS, Ph.D Yale 1977 The regional distribution of federal taxes and expenditures· A study in the theory and estimation of fiscal incidence.

ROBERT D. CHRISTIE, Ph.D. Queen's (Kingston) 1978. Money, inventories, and the stability of full-employment equilibrium.

MAY P. CHU, Ph.D. Case Western Reserve 1978. Substitution among bank loans, commercial paper, and certificates of deposit.

ROBERT CLINE, Ph D Michigan 1977. An econometric study of state budgeting: The Michigan general fund-general purpose budget

PARAMJEET K. DHALIWAL, Ph.D Kentucky 1976. The effect of size on cost of credit unions in Kentucky.

LYNNE M DOTI, Ph.D. California (Riverside) 1978 Banking in California· Some evidence on structure, 1878 1905.

BARRY B. DOUGLAS, Ph.D California (Davis) 1977 An empirical and theoretical application of game theory to municipal expenditure decisions.

LARRY V. ELLIS, Ph.D. Missouri (Columbia) 1978 Wealth, government debt, and the crowding out of fiscal policy actions.

JERRY T FERGUSON, Ph D Florida 1978 A study of the impact of the use value property taxation program in Virginia.

MARK J FLANNERY, Ph.D. Yale 1978. Financial intermediation under uncertainty: A micro-economic analysis.

CLYDE A. GARNER, Ph.D. Harvard 1977 Capital flows and Canadian monetary policy: An empirical study.

THOMAS H. GALE, Ph.D. Wisconsin (Madison) 1977 The determinants of public expenditures in rural Wisconsin communities.

ROBERT GENTENAAR, Ph.D. Michigan State 1977. The consumer sector's demand for assets and the supply of corporate bonds and equities: An integration of portfolio theory and a theory of the firm.

FRANK W. GIESBER, Ph.D. Texas (Austin) 1978. Direct protection of ultimate consumers by the government of the state of Texas as economic and legal analysis.

MICHAEL L. GOETZ, Ph.D. Minnesota 1977. Tax evasion as a determinant of the optimal level of tax collection expenditures.

DERRICK K. GONDWE, Ph.D. Manitoba 1978. The incidence and economic effects of indirect taxation in Malawi.

JOSHUA GREENE, Ph.D. Michigan 1977. The economics of making government an employer of last resort in the United States: Some important issues.

SHAWNA GROSSKOPF, Ph.D. Syracuse 1977. Discrimination in state and local government employment.

JOHN H. HAMMOND, Jr., Ph.D. George Washington 1977. Fixed investment models and expectations. Implications of a new data source.

ROGER P HARMAN, Ph.D. Pennsylvania 1977. Measuring local fiscal capacity.

JOHN J. HARRINGTON, JR., Ph.D. New York 1977 An investigation of the effects of disintermediation and reintermediation upon the deposit growth and mortgage holdings of mutual savings banks and savings and loan associations.

CHANG-TSEH HSIEH, Ph.D. Purdue 1978. The financial behavior of households with special emphasis on the supply of mortgages and demand for passbook and certificate accounts.

ROBERT L. JONES, Ph D Notre Dame 1978 The implications of commercial bank liability management upon monetary policy

BRENDA J. KAHN, Ph D. Columbia 1978 Expected rate of change in prices and foreign exchange rates under a dual money system

WILLIAM R. KEETON, Ph.D Massachusetts Institute of Technology 1977 Equilibrium credit rationing.

LAURENCE J. KOTLIKOFF, Ph D. Harvard 1977 Essays on capital formation and social security, bequest formation, and long-run tax incidence.

BYUNG-SUB KWAK, Ph D City (New York) 1978 A study of government revenue from money creation in South Korea.

RICHARD W. LANG, Ph D Ohio State 1977. Coupon bonds, duration, and liquidity premia. A study of the liquidity-preference theory of the term structure of interest rates

SUNG WHI LEE, Ph.D. Columbia 1978. Estimation of liquidity in a macro-economic model and its comparative performance with conventional definitions of money

SEE-YAN LIN, Ph.D. Harvard 1977. Malaysia: Money and monetary management, 1957–76.

ROLAND LIPKA, Ph.D. Rutgers 1977. Some extensions to the Harberger tax incidence model: A simulation study.

ROBERT T. MCGEE, Ph.D. Wisconsin (Madison) (1978). The impact of Federal Reserve behavior on the dynamic structure of the money supply mechanism.

DOROTHY M. MERCER, Ph.D. Oklahoma State 1977. An economic analysis of the technical, allocative, and equity effects of financing municipal government from revenue over cost earned by the municipal electric utility: A case study of Stillwater, Oklahoma.

ANNETTE MEYER, Ph.D. City (New York) 1978. Comparative study of changes in the budgetary process of France and the United States: 1921 to 1971.

ELLIOTT MIDDLETON III, Ph.D. Colorado 1978. The plunging catastrophe. A model of discontinuous portfolio adjustment.

RANDALL MILLER, Ph.D. Pittsburgh 1977. A theoretical and empirical investigation into the regional impact of monetary policy in the United States.

FREDERICK S. MISHKIN, Ph.D. Massachusetts Institute of Technology 1977. Illiquidity, the demand for consumer durables, and monetary policy.

DOUGLAS W. MITCHELL, Ph.D. Princeton 1978. Interest-bearing checking accounts and macro policy

STEPHEN D. NADAULD, Ph.D. California (Berkeley) 1978. The interest elasticity of net worth in savings institutions.

JOSEPH P. O'BRIEN, Ph.D. Oklahoma State 1977 Federal Reserve policies and variability of the money supply.

CAROL C. O'CLEIREACAIN, Ph.D. London School of Economics The determinants of local government expenditure in English and Welsh county boroughs, 1971.

JOHN S OH, Ph.D. Virginia 1978. The long- and short-run demand for money in Australia: 1952 to 1970.

MANOON PAHIRAH, Ph.D. Hawaii 1978. Tax incidence: A case study of Thailand

ANTHONY J. PELLECHIO, Ph D. Harvard 1978. Social Security and retirement behavior.

AVRAHAM PONIACHEK, Ph.D. State University of New York (Albany) 1977. Monetary independence under flexible exchange rates: The theoretical issues and the recent West German experience.

JOEL L. PRAKKEN, Ph.D. Washington (St. Louis) 1977. The stability of the interest elasticity of the demand for money.

ANIL K PURI, Ph D. Minnesota 1977. Tax structure change: A theory and empirical evidence.

DONALD L. RAIFF, Ph.D. Ohio State 1978. The effects of Federal Reserve System's operations on the supply of legal reserves 1890-1935.

MUHAMMAD RASHID, Ph.D. Queen's (Kingston) 1978. The theory and measurement of the cost of capital to the Canadian economy.

MARTIN REGALIA, Ph.D. Wisconsin (Madison) 1977.

A test of the assumption of exogenous reserve targets in small money supply models.

WILLIAM R. REICHENSTEIN, Ph.D. Notre Dame 1978. The definition of money: The implications of alternative methodologies in estimating distributive lags.

ARNOLD P. REZNEK, Ph.D. Iowa 1978. An econometric model of the government sector of the state of Iowa.

V. VANCE ROLEY, Ph.D. Harvard 1977. A structural model of the *U.S.* government securities market.

DAVID M. ROWE, Ph.D. Pennsylvania 1977. Financial flows in nonfinancial business.

BRUCE L. RUBIN, Ph.D. Case Western Reserve 1978. An analysis of the price behavior of closed-end investment companies.

ETHAN SEIDEL, Ph.D. Johns Hopkins 1978. The effect of bank credit cards and use on the demand for money by the household sector.

MOHAMMAD R. SHAHROODI, Ph.D. Syracuse 1978. Iranian budgetary behavior, 1960–76.

FREDERICA SHOCKLEY, Ph.D. Georgia State 1978. The incidence of the property tax on residential rental property in Orange and Seminole counties.

EUGENIE D. SHORT, Ph.D. Virginia 1978. A theoretical and empirical analysis of the productivity of money

KENNETH SINGLETON, Ph.D. Wisconsin (Madison) 1977. The cyclical behavior of the term structure of interest rates.

ENID SLACK, Ph.D Toronto 1978. The budgetary response of municipal governments to provincial transfers: The case of Ontario.

JEFFREY L. SMITH, Ph.D. Chicago 1977. Some evidence on the relationship between politics and inflation.

PETER SPERLING, Ph.D. City (New York) 1978. Six essays on the government revenues and welfare losses from money creation.

THOMAS J. STEFFANCI, Ph.D. Connecticut 1978. The impact of size and market structure on commercial bank portfolio substitution and adjustment.

RAY G. STEPHENS, D.B.A. Harvard 1978. Uses of financial information in structuring and improving decision processes for bank lending officers.

RICHARD THALHEIMER, Ph.D. Kentucky 1976. A portfolio choice model with an application to mutual savings banks.

JOHN M. TOMA, Ph.D. Virginia Polytechnic Institute 1977. Institutional structures and local government consolidation.

JOHN TRIANTIS, Ph.D. New Hampshire 1978. An evaluation of the leading economic indicators in the monetary transmission process.

PAUL A. VOLKER, Ph.D. Simon Fraser 1977. Aspects of the effects of direct controls on the Australian postwar monetary sector.

NANCY WENTZLER, Ph.D. Wisconsin (Madison) 1977. Equality of opportunity and the distribution of state aid to school districts.

JOHN A. WEYMARK, Ph.D. Pennsylvania 1977. Essays in public economics.

ELLIOTT S. WILLMAN, Ph.D. Indiana 1977. The effectiveness of monetary policy in an open economy under fixed exchange rates: The case of Germany and the United Kingdom.

STEVEN WOODS, Ph.D. Claremont 1977. An appraisal of Federal Reserve actions, 1953–75.

LOUIS K. W. YING, Ph.D. Purdue 1977. Theories of consumer credit.

International Economics

JOSE L ALBERRO, Ph.D. Chicago 1978. Essays on the open monetary economy.

SUSAN L ALEXANDER, Ph.D. Southern Methodist 1977. Pure intermediate goods in the theory of protection

MOISE ALLAL, Ph.D. Michigan State 1977. Evaluation of the African Associated States response to tariff preferences granted by the European Economic Community

PETER F. ALLGEIER, Ph.D. North Carolina (Chapel Hill) 1977. Economic sukiyaki: A multitheory approach to *U S* -Japanese trade.

KYM ANDERSON, Ph.D Stanford 1977 Distributional aspects of trade protection in Australia, with emphasis on the rural sector.

GOSAH ARYA, Ph.D. Pittsburgh 1976. A structural model of the Thai balance of payments: 1956–73.

ATA A. ATMAR, Ph.D. Clark 1977. The role of the cartel in the new international order.

PAT BALAN, Ph.D. State University of New York (Binghamton) 1978. The cost of accepting economic aid: A case study of Soviet aid to India.

MANUEL S. BARBOSA, Ph.D. Yale 1977. Growth, migration, and the balance of payments in a small open economy.

KENNETH BERCUSON, Ph.D. Yale 1977 Capital movements, portfolio balance, and internal equilibrium in a small open economy.

CAROLYN L. K. BOMBERGER, Ph.D. Brown 1978. A portfolio approach to international payments adjustment: The case of a large economy.

ANTONIO P. BRANDAO, Ph.D. Purdue 1978. New perspectives on the terms of trade and the gains from trade: A case study of Brazil.

STEVEN R. BRENNER, Ph.D. Stanford 1977. Economic interests and the trade agreements program, 1934–40: A study of institutions and political influence.

ROBERT BRUSCA, Ph.D. Michigan State 1977. An empirical examination of several theories of the commodity composition of trade.

HENRY E. BWAMBALE, D.B.A. Harvard 1978. Agricultural research and technology: Diffusion by foreign agribusiness firms in Kenya.

HUSSEIN K. CHALABI, Ph.D. Columbia 1978. Expectations and exchange market efficiency: An empirical investigation.

WING TO CHAN, Ph.D. Rochester 1978. The effect of urban minimum wage on unemployment, capital mobility, and international trade.

DONALD V. COES, Ph.D. Princeton 1978. Exchange rate uncertainty and the structure of Brazilian foreign trade.

KEITH J. COLLINS, Ph.D. North Carolina State 1977. An economic analysis of export competition in the world coarse grain market: A short-run constant elasticity of substitution approach.

JOHN CUDDINGTON, Ph.D. Wisconsin (Madison) 1978. An integration of economic growth models in pure trade theory and international finance.

ARDESHIR J. DALAL, Ph.D. Iowa 1977. Decision making in forward exchange markets.

SATYA P. DAS, Ph.D. Southern Methodist 1977. Analysis of devaluation in the presence of international investment.

JOSE D. DASILVEIRA, Ph.D. Florida 1977. A simple model of exchange market pressure applied to the Brazilian case.

MARIO DRAGHI, Ph.D. Massachusetts Institute of Technology 1977. Essays on economic theory and applications.

ROBERT DRISKILL, Ph.D. Johns Hopkins 1978. Monetary policy and exchange rate dynamics.

LIAM P. EBRILL, Ph.D. Harvard 1977. Elements of a Keynesian approach to floating exchange rates.

HAMID ETEMAD, Ph.D. California (Berkeley) 1978. A game-theoretic approach to the host country multinational corporation relations.

RICHARD A. FEY, Ph.D. Brown 1978. A contribution to the theory and econometrics of international trade in manufacturers: The case of more goods than factors of production.

DENNIS O. FLYNN, Ph.D. Utah 1978. The Spanish price revolution and the monetary approach to the balance of payments.

MARCOS G. DA FONSECA, Ph.D. Yale 1978. The general equilibrium effects of international trade policies.

PAUL GERNANT, Ph.D. Michigan 1977. The international trade effects of the 1965 *U.S.*-Canadian automotive agreement.

ALAN G. GOEDDE, Ph.D. Duke 1978. *U.S.* multinational manufacturing firms: The determinants and effects of foreign investment.

DENNIS E. GOODMAN, Ph.D. Southern Illinois 1977. Innovation and international trade.

PAULO R. GUEDES, Ph.D. Chicago 1978. Fiscal policy, public debt and external indebtedness in nonmonetary two-sector open growth models.

ABOLGHASEM HASHEMI, Ph.D. Indiana 1978. Technological disparities and tariff protection in industrial countries.

JEAN F. HENNART, Ph.D. Maryland 1977. A theory of foreign direct investment.

DIT S. HO, Ph.D. Minnesota 1977. An econometric study of the relationship between international and domestic prices in the Japanese economy.

TAKAO ITAGAKI, Ph.D. Southern Methodist 1977. Theory of multinational firms under fixed and flexible exchange rates.

ELIOT R. J. KALTER, Ph.D. Pennsylvania 1978. Effect of exchange rate changes upon matched domestic and export prices.

PETER M. KELLER, Ph.D. Rochester 1978. Cyclical migration, employment, and payments with some application to the German economy.

RICHARD F. KENNEDY, Ph.D. Rice. A study of French trade flows, 1966–76.

FRANK F. KIANG, Ph.D. Oklahoma State 1977. The feasibility of mutually beneficial trade negotiations between the United States and its major trading partners among less developed countries.

HAK SU KIM, Ph.D. South Carolina 1977. Monetary policy consideration of the foreign capital inflows in a dependent economy: The case of Korea.

SOO-YONG KIM, Ph.D. Michigan State 1978. Transmission of international economics fluctuations to a small open economy: The case of Korea.

BOHN-YOUNG KOO, Ph.D. George Washington 1977. Industrial characteristics and patterns of trade: A test of alternative trade theories on the Korean manufacturing industries.

PRAIPHOL KOOMSUP, Ph.D. Yale 1978. Export instability and export diversification: A case study of Thailand.

PANAYOTIS E. KOVEOS, Ph.D. Pennsylvania 1977. The transition to flexible exchange rates and international portfolios.

IRVING H. KUCZYNSKI, D.B.A. Harvard 1978. British offshore oil and gas policy.

JEAN-FRANCOIS LANDEAU, D.B.A. Harvard 1977. The international stratagies of the *U.S.* independent oil companies.

SOYNO LEE, Ph.D. Northern Illinois 1977. Trade and economic growth of South Korea: 1955–75.

LEONARDO LEIDERMAN, Ph.D. Chicago 1978. Expectations, output-inflation tradeoffs, and the balance of payments in a fixed exchange rate economy.

JAMES MCKEE, Ph.D. Syracuse 1977. A simulation of an international monetary institution.

MICHAEL R. MCMAHON, Ph.D. Washington 1978. The

capital account approach to international monetary analysis.

JOHN MOONEY, Ph.D. California (Los Angeles) 1978. Short-run determinants of exchange movements: The *U.S.* dollar/German mark rate (1974–75).

JOANNA MOSS, Ph.D. New School 1978. The Yaounde Convention, 1964–75.

ALKIMAR R. MOURA, Ph.D. Stanford 1978. Private external borrowing: The Brazilian experience.

JOHN D. MURRAY, Ph.D. Princeton 1978. Tax differentials and international capital flows: The Canadian-*U.S.* experience.

TEVFIK F. NAS, Ph.D. Florida State 1978. Effects of Turkish-*EEC* customs union on Turkish manufactured exports.

RUEDIGER NAUMANN-ETIENNE, Ph.D. Michigan 1977. Exchange risk in foreign operations of multinational corporations.

ASGER M. NIELSEN, Ph.D. Michigan 1977. Economic impact audit: A case study approach to measurement of costs and benefits to a host state in the United States from reverse foreign direct investment.

KAREN J. OHLDIECK, Ph.D. Florida 1977 Money neutralization and the balance of payments of Norway.

LOUKA K. PAPAEFSTRATIOU, Ph.D. Princeton 1978. Transmission of external price disturbances in small open economies.

M. RAY PERRYMAN, Ph.D. Rice 1978. An indicator of monetary policy derived from a simultaneous equation model.

STEVEN E. PLAUT, Ph.D. Princeton 1978. A treatise on vulnerability or price changes of imported intermediate goods and the welfare of the importing country.

PEDRO POU, Ph.D. Chicago 1978. Money and the balance of payments: The experience of Argentina and Brazil.

JOHN PRICE, Ph.D. California (Los Angeles) 1978. Time-series analysis of money supply reaction functions: International evidence.

CHANPEN PUCKAHTIKOM, Ph D. Rochester 1978. Balance of payments and monetary developments: Thailand, 1947–73.

THOMAS A. PUGEL, Ph.D. Harvard 1978. The effect of international market linkages on prices, profits, and wages in *U S.* manufacturing industries.

ALAN A. RABIN, Ph.D. Virginia 1977. A monetary view of the acceleration of world trade inflation, 1973–74.

HARRY RAMCHARRAN, Ph.D. State University of New York (Binghamton) 1978. Trade creation and trade division in the Caribbean free trade area: An empirical study.

ALLEN ROSS, Ph.D. Columbia 1978. Human capital and technology in international trade.

SUKRITA SACHCHAMARGA, Ph.D. Southern Methodist 1978. The effect of exchange rate changes on direct international investment.

ALEXANDER H. SARRIS, Ph.D. Massachusetts Institute of Technology 1977. The economics of international grain reserves systems.

CHI C. SCHIVE, Ph.D. Case Western Reserve 1978. Direct foreign investment, technology transfer, and linkage effects: A case study of Taiwan.

JAMES R. SCHMIDT, Ph.D. Rice 1978. Economically rational expectations and the demand for money.

KOON-LAM SHEA, Ph.D. Washington (St. Louis) 1977. Imported inputs and devaluation.

SHUN-YI SHEI, Ph.D. Purdue 1978. The exchange rate and *U.S.* agricultural product markets: A general equilibrium approach.

BRUCE SMITH, Ph.D. Columbia 1978. Purchasing power parity, the real exchange rate, and the theory of flexible exchange rates.

ALAN STOCKMAN, Ph.D. Chicago 1978. A theory of exchange rate determination.

PEGGY S. SWANSON, Ph D. Southern Methodist 1978. Foreign holders of short-term liquid dollar assets

LYDIA THACKREY, Ph.D Pittsburgh 1977. A simultaneous equation approach to *U S* commodity exports.

PRANEE TINAKORN, Ph D Pennsylvania 1978. An empirical evaluation of the *UNCTAD* integrated commodity program.

ERNESTO TIRONI, Ph.D. Massachusetts Institute of Technology 1977 Economic integration and foreign direct investment policies· The Andean case.

GIUSEPPE TULLIO, Ph.D. Chicago 1977. Monetary equilibrium and balance-of-payment adjustment: An empirical test of the *U.S.* balance of payments.

RICARDO VARSANO, Ph.D. Stanford 1977. Border tax adjustments, factor mobility, and growth.

ROBERT G WILLIAMS, Ph.D. Stanford 1978. The Central American common market: Unequal benefits and unequal development

YOSHINORI YOKOI, Ph.D. Colorado 1978. Studies in the monetary approach to international finance.

ISAO YOROZU, Ph.D. Claremont 1978. Japanese-American trade relations· An empirical application of input-output analysis.

EMMANUEL J. ZERVOUDAKIS, Ph.D. Rochester 1978. Determinants of the dollar-sterling rate, 1919–25, and some related issues.

Business Administration; including Business Finance and Investment, Insurance, Marketing, and Accounting

RONY M. ADELSMAN, Ph.D. Purdue 1977. Analysis of group decision-making models under varying information conditions and assumptions regarding individual behavior.

Gerald B. Allan, D.B.A. Harvard 1978. Competitive behavior and corporate growth.

George Applewhite, Ph.D. New York 1977. Investigation of the nature of announced and unaudited annual earnings.

Roger C. Bennett, D.B.A. Harvard 1978. Centralized purchasing in multidivisional companies.

William R. Boulton, D.B.A. Harvard 1977. The nature and format of director information flows: An exploratory study.

Marinus J. Bouwman, Ph.D. Carnegie-Mellon 1978. Financial diagnosis: A cognitive model of the processes involved.

Lance M. Brofman, Ph.D. New York 1978. Implications and applications of organized option markets for financial theory.

Jorge R. Calderon-Rossell, Ph.D. Michigan 1978. A multinational firm sourcing model.

Richard P. Castanias II, Ph.D. Carnegie-Mellon 1978. Essays on the behavior of asset prices under uncertainty.

Lewis Chakrin, Ph.D. New York 1978. Investment choice, market equilibrium, and corporate finance in a state-dependent utility framework.

Guy Charest, Ph.D. Chicago 1978. Split in dividend information, stock returns, and market efficiency.

Joseph L. Cheng, Ph.D. Michigan 1977. Organizational coordination, integration, interdependence, and their relevance to research unit effectiveness: A comparative study.

Joseph K. Cheung, Ph.D. Michigan 1977. The predictive relevance of accounting numbers.

Roberta N. Clarke, D.B.A. Harvard 1978. A study of the revitalization of mass packaged goods.

Martin N. De Waele, Ph.D. California (Berkeley) 1978. The design of decision aids for marketing managers, based on Jungian personality types and other managerial styles.

Gadis J. Dillon, Ph.D. Michigan 1977. The role of accounting in the stock market crash of 1929.

Lowell Dworin, Ph.D. Michigan 1977. Inflation and corporate taxation: An economic analysis.

Douglas B. Engel, Ph.D. Michigan 1978. Functional coupling, environmental coupling, and organizational structure in technological innovation.

Mario C. Ferrario, D.B.A. Harvard 1978. Strategic management in state enterprises.

John K. Ford, D.B.A. Harvard 1977. Integration of the theories of production and finance.

Anna C. Fowler, Ph.D. Texas (Austin) 1977. Charitable remainder trusts under the post-1969 tax law: Effects of changes in the law on their use, a model for tax planning, and suggestions for revisions.

Robert N. Freeman, Ph.D. Texas (Austin) 1977. Accounting and market estimates of risk due to unanticipated inflation.

Merle E. Frey, Ph.D. New York 1977. Effects of task characteristics and individual differences on performance-job satisfaction relationships.

Haim D. Fried, Ph.D. New York 1978. An examination of the aggregation problem in accounting under the predictive ability criterion.

George W. Gallinger, Ph.D. Purdue 1977. An industrial organization explanation of corporate cash tender offers.

John E. Gilster, Ph.D. Michigan 1977. Efficient frontier composition: The holding period problem.

Thomas W. Hackett, Ph.D. Oregon 1978. A simulation analysis of immunization strategies applied to bond portfolios.

Dominique M. Hanssens, Ph.D. Purdue 1977. An empirical study of time-series analysis in marketing model building.

Jane L. Holtz, D.B.A. Harvard 1978. Exploring the psychological contract over the life cycle.

William S. Hopwood, Ph.D. Florida 1978. An empirical investigation into the usefulness of accounting and market data in the forecasting of accounting earnings.

Linda Jewell, Ph.D. Florida 1978. Leadership and the group-induced shift: A field study of a complex decision problem.

Thomas M. Jones, Ph.D California (Berkeley) 1977. Shareholder suits: A contemporary survey of their utilization.

Lawrence I. Kessler, Ph.D. Texas (Austin) 1977. The effect of different types of cognitive feedback upon the prediction achievement of accounting users: Some experimental evidence.

Joseph Liberman, Ph.D. Chicago 1977. Human capital and the financial capital market: An empirical investigation.

Roe M. Lobue, Ph.D. Florida State 1978. The return-risk properties of bank holding company acquisition.

Thomas V. McCullough, Ph.D. California (Berkeley) 1977. A campus planning model.

Hassell H. McClellan, D.B.A. Harvard 1978. One-bank holding companies: A study of management.

Vinay V. Marathe, Ph.D. California (Berkeley) 1978. Elements of covariance in security returns and their macro-economic determinants.

Ronald W. Masulis, Ph.D. Chicago 1978. The effects of capital structure change on security prices.

Alan D. Meyer, Ph.D. California (Berkeley) 1977. Hospital environment, strategy, and structure: The role of managerial perception and choice.

Eitan Muller, Ph.D. Northwestern 1977. Information, persuasion, and advertising policies.

Albert E. Muir, Ph.D. State University of New York (Albany) 1977. The determinants of extramural research support to university and college faculty.

OVE K. MYRSETH, D.B.A. Harvard 1978. Intrafirm diffusion of organizational innovations: An exploratory study.

DONALD P. NEWMAN, Ph.D. Texas (Austin) 1977. Toward an integrative theory of accounting information production and utilization.

PETER J. O'CONNOR, Ph.D. Florida 1978. The use of brand preference measures for market segmentation.

BRIAN A. O'DOHERTY, Ph.D. Florida 1978. Modern accounting approaches to capital maintenance and valuation.

EDWARD T. POPPER, D.B.A. Harvard 1978. Structural and situational effects on mother's response to children's purchase requests.

SAMUEL RABINO, Ph.D. New York 1978. *DISC*'s contribution to improved competitiveness in export markets: A study of the perceptions of American manufacturing exporters.

JOEL C. RENTZLER, Ph.D. New York 1978. Pricing generalized European options in complete markets and variable parameter regression in *Beta* analysis.

MARSHALL B. ROMNEY, Ph.D. Texas (Austin) 1977. The elicitation of auditors' prior probability distributions for variables estimation.

LOUIS L. ROQUET, D.B.A. Harvard 1978. The process of top management succession in large public companies.

RONALD S. ROSS, Ph.D. Texas (Austin) 1978. An analysis of *DISC* legislation and comparison with export incentives of selected foreign countries.

NORBERT V. SCHAEFER, Ph.D. California (Berkeley) 1977. The change in business ideology.

BARRY B. SCHWEIG, Ph.D. Pennsylvania 1977. An analysis of the effectiveness of product liability underwriters.

JOHN A. SEEGER, D.B.A Harvard 1978. Changing problem solving behavior in management meetings: Will the meeting please come to order?

J. DOUGLAS SHOLUND, Ph.D. Purdue 1977. An investigation of investor expectations and security valuation.

DANIEL G. SHORT, Ph.D. Michigan 1977. The usefulness of price-level adjusted accounting numbers in the context of risk assessment.

STEVEN M. SHUGAN, Ph.D. Northwestern 1978. Descriptive stochastic preference theory and dynamic optimization: Applications toward predicting consumer choice among finite alternatives with marketing implications.

RONALD W. SKEDDLE, Ph.D. Case Western Reserve 1977. Empirical perspectives on major capital decisions.

LUC A. SOENEN, D.B.A. Harvard 1977. Foreign exchange exposure management for international business firms: A portfolio approach.

BRUCE N. STRAM, Ph.D. Maryland 1976. An analysis of treatment processing production systems.

STEPHEN A. STUMPF, Ph.D. New York 1978. Identifying optimal groups for making judgmental decisions: An experimental study of metadecision making.

NATALIE T. TAYLOR, D.B.A. Harvard 1978. Management succession under conditions of crisis: The role of the board of directors.

ANDREW F. THOMPSON, Ph.D. Nebraska (Lincoln) 1977. An analysis of the policy loan interest rate as a decision variable in granting nonforfeiture benefits on ordinary life insurance contracts.

GERTRUDE G. VERSER, D.B.A. Harvard 1978. The effects of an interfunctional power imbalance on new product development in marketing-oriented firms.

WANDA A. WALLACE, Ph.D. Florida 1978. The impact of selected financial reporting practices and the nature of the audit opinion upon municipal interest cost and bond rating.

MARK I. WEINSTEIN, Ph.D. Chicago 1977. An examination of the behavior of corporate bond prices.

WARREN R. WILHELM, D.B.A. Harvard 1977. A study of the lives of twenty organizational managers aged thirty-five to forty-five.

MARK A. WOLFSON, Ph.D. Texas (Austin) 1977. Toward an understanding of the complementary nature of security price and nonsecurity price information in relative risk parameter estimation.

Industrial Organization and Public Policy; including Economics of Technological Change, and Industry Studies

JACK ALLENTUCK, Ph.D. New School 1978. Innovation and the structure of the electric utility industry.

HENRY O. ARMOUR, Ph.D. Stanford 1977. The *U.S.* petroleum industry Vertical integration, organizational structure, and economic performance.

CHRISTOPHER BARNEKOV, Ph.D. Chicago 1978. The impact of market structure on performance in domestic airline markets.

JACK E. BARRAR, Ph.D. Oregon 1978. A multinomial logit model of the journey to school: A study of individual travel demand.

DONALD BASCH, Ph.D. Yale 1977. Response to regulatory change: The case of *NOW* accounts in Massachusetts, 1972–75.

STEPHEN G. BUELL, Ph.D. Lehigh 1977. The relationship between earnings retension and subsequent growth in earnings per share for large American companies, 1948–68.

JON H. BURKMAN, Ph.D. Pittsburgh 1977. Technological change and productivity growth in the textile industries of Japan and Hong Kong.

EROL CAGLARCAN, Ph.D. George Washington 1977. The economics of innovation in the pharmaceutical industry.

DEXTER J. L. CHOY, Ph.D. Hawaii 1978. Airline pricing of domestic passenger charters.

MICHAEL COHN, Ph.D. New York 1977. The demand for automobiles in the suburban counties of the New York *SMSA*.

GRENVILLE CRAIG, Ph.D. Yale 1977. Perspectives on motivation, barriers, and consequences.

THI DUYDAN DAO, Ph.D. Pittsburgh 1976. Economic effects of durability improvement: A case study of new automobiles.

PETER J. DUNNETT, Ph.D. Simon Fraser 1977. The effect of British government policy on the British motor industry, 1945–75.

IBRAHIM S. ELRIFADI, Ph.D. Pittsburgh 1976. A study in the economics of the Libyan oil sector.

PHILIP G. FAVERO, Ph.D. Michigan State 1977. The processes of collective action: Small electric companies in Michigan.

RICHARD L. FERGUSON, Ph.D. Stanford 1978. An economic analysis of the Securities Act of 1933.

LOUIS A. FERLEGER, Ph.D. Temple 1978. Technological change in the postbellum Louisiana sugar industry.

JOHN FILER, Ph.D. Chicago 1977. An economic theory of voter turnout.

THERESA A. FLAIM, Ph.D. Cornell 1977. The structure of the *U.S.* petroleum industry. Concentration, vertical integration, and joint activities.

SAM J. FRASER, Ph.D Louisiana State 1978. Political economic cartels: An alternative approach to the world oil market.

H. LANDIS GABEL, Ph.D. Pennsylvania 1977. A simultaneous equation analysis of industrial structure and performance.

DANIEL J. GALLAGHER, Ph.D. Maryland 1976. An economic analysis of the player reservation system in the professional team sports industry in the United States

GEORGE B. GARMAN, Ph.D. Notre Dame 1978. A study of the process in the American steel industry using the translog production function.

FREDERICK H. GAUTSCHI III, Ph.D. California (Berkeley) 1978. Adjudicative decisions in the Federal Trade Commission.

RANDALL R. GEEHAN, Ph.D. Massachusetts Institute of Technology 1977. The production of financial and insurance services.

DAVID GLASNER, Ph.D. California (Los Angeles) 1977. The effects of rate regulation on automobile insurance premiums.

ANTHONY J. GRECO, Ph.D Tennessee (Knoxville) 1978. State regulation of fluid milk and the processor-retailer margin.

MAGNI GUDMUNDSSON, Ph.D. Manitoba 1977. The Danish monopolies legislation.

VEENA A. GUPTA, Ph.D. Tufts 1978. A study of price and cost behavior of the chemicals industry in the United States.

VINOD K. GUPTA, Ph.D. Toronto 1977. Structure, conduct, and performance in Canadian manufacturing industries: A simultaneous equations approach.

CHRISTOPHER D. HALL, Ph.D. Washington 1978. Essays in the costs of transacting ideas.

WILLIAM S. HALLIGAN, Ph.D. California (Davis) 1977. Theories of sharecropping and share contracting for California gold.

H. MICHAEL HAYES, Ph.D. Michigan 1977. The effective salesman of electric distribution equipment as perceived by buyers in the electric utility industry.

KENNETH H. HELLER, Ph.D. Texas (Austin) 1977. The impact of *U.S.* income taxation on the financing and earnings remittance decisions of *U.S.* based multinational firms with controlled foreign corporations.

IBRAHIM A.-S. IBRAHIM, Ph.D. Duke 1977. Econometric analysis of the performing and visual arts.

FRANK O. IRVINE, Ph.D. Massachusetts Institute of Technology 1977. A study of markets and the inventory and pricing policies of market-maker firms.

EDWIN K. ISLEY, Ph.D. Notre Dame 1977. The effect of user charges on the demand for barge transportation of grain.

GREGG A. JARRELL, Ph.D. Chicago 1978. The demand for and effects of state regulation of the electric utility industry.

STEPHEN B. JARRELL, Ph.D. Purdue 1978. Research and development and firm size in the pharmaceutical industry.

RONALD N. JOHNSON, Ph.D. Washington 1977. Competitive bidding for federally owned timber.

FREDERICK W. JONES, Ph.D. Virginia 1978. Input biases under rate of return regulation: A test of an intertemporal Averch-Johnson model.

WILLIAM J. JORDAN, Ph.D. State University of New York (Albany) 1977. The optimal pricing structure for the natural gas pipeline industry.

FRANK P. JOZSA, Ph.D. Georgia State 1977. An economic analysis of franchise relocation and league expansion in professional team sports, 1950–75.

JOHN JUREWITZ, Ph.D. Wisconsin (Madison) 1978. The internalization of environmental costs in the private electric utility industry.

NYLE KARDATZE, Ph.D. California (Los Angeles) 1978. Intercity differences in retail gasoline price fluctuations.

FARHAD KHAMSI, Ph.D. New School 1977. The classification of technological change.

KENNETH KOFORD, Ph.D. California (Los Angeles) 1977. An economic theory of legislatures: Centralized exchange under competition and monopoly.

OTTO G. KONZEN, Ph.D. Wisconsin (Madison) 1977. Effects of a program to increase yields on farm organization and income: A longitudinal analysis of Brazilian farms, 1969–73.

GEORGE KROON, Ph.D. California (Los Angeles) 1978. Advertising, innovation, and profit.

STANLEY LANCEY, Ph.D. Pennsylvania 1978. A quarterly econometric model of the *U.S.* wood products industry.

EDWARD J. LINCOLN, Ph.D. Yale 1978. Technical change on the Japanese national railways, 1949–74.

DAVID S. LINDSAY, Ph.D. California (Los Angeles) 1978. Pricing tactics, regulation, and successive monopoly.

JOHN J. LOMBARD, Ph.D. Connecticut 1978. Electric utility rate design: An application of peak load pricing to a major electric utility in New England

CHARLIE A. MCCORMICK, Ph.D Virginia Polytechnic Institute 1978. Intrastate telephone regulation: A public choice approach.

EDGAR J. MCDOUGALL, JR., Ph.D. Florida 1978. A consumer utility model for the allocation of sales among major centers.

HENRY B. MCFARLAND, Ph.D. Northwestern 1978. The estimation of railroad cost functions.

DOUGLAS W. MCNEIL, Ph.D. Oklahoma State 1977. The economic impact of mechanically deboning red meats.

NITIN T. MEHTA, D.B A. Harvard 1978. Policy information in a declining industry The case of the Canadian dissolving pulp industry.

JOHN B. MEISEL, Ph.D. Boston College 1978. The determinants of advertising expenditures: Theory and evidence.

THOMAS NAGLE, Ph.D. California (Los Angeles) 1978. Consumer information, advertising, and industry profitability theory and verification

R. D. NAIR, Ph.D. Michigan 1977. The sensitivity of tests of corporate investment models to accounting methods of inventory valuation and depreciation An empirical analysis.

DAVID J. NICOL, Ph.D. Case Western Reserve 1977. Economies of scale in the production of air transportation.

DOUGLAS L. NORLAND, D.B.A. Indiana 1977. An *ex ante* measurement of the Averch-Johnson effect in electric power generation.

KENNETH R. NOWOTNY, Ph.D. Texas (Austin) 1977. An empirical investigation of the "Averch-Johnson Hypothesis" as it applies to the telephone subsidiaries of the American Telephone & Telegraph Company.

JOHN M. PATRICK, Ph.D. Michigan State 1977 An economic analysis of improving the variability of rail lines: A Michigan case study.

ROBERT C. PEBWORTH, JR., Ph.D. Indiana 1978. The Ruhr Coal Cartel: An economic analysis of an organized market.

JOHN K. PIERCEY, Ph.D. Oklahoma 1978. The pipeline segment of the domestic petroleum industry: Structure and conduct.

KATHLEEN A. PULLING, Ph.D. California (Riverside) 1977. Market structure and the cyclical behavior of prices and profits, 1949–75.

HOWARD L. REESE, Ph.D. Kansas 1978. Regional firm size differentials in antebellum *U.S* manufacturing.

ROLAND L. REYNOLDS, Ph.D. Washington State 1977. Interrelationships between market structure and pollution abatement activities.

ANIS B. SALIB, Ph.D. Vanderbilt 1977. An economic analysis of international differences in energy use: An input-output approach.

KATHERINE A. SCHIPPER, Ph.D. Chicago 1977. An analysis of the financial condition of private colleges.

SERGIO S. SEPULVEDA, Ph.D. Cornell 1978. The impact of modern technologies upon employment and factors' return in integrated rural development districts in Columbia.

PHILLIP S SHINODA, Ph.D California (Berkeley) 1978. Conglomerators and the press.

EVANGELOS SIMOS, Ph.D Northern Illinois 1977 Capital utilization, its measurement and its implications for production function and economic growth.

ANTONIO D. SOBRINHO, Ph.D. Cornell 1978. Technology and performance of Brazilian and foreign firms in San Paulo

JAMES SONDA, Ph.D. Michigan 1977. Technology forcing and auto emissions control.

MAX H. STRADER, Ph D. Florida 1977. A cross-section analysis of the demand for mobile homes

JAMES B. SUMRALL, JR., Ph.D. Boston College 1978. The diffusion of the basic oxygen furnace in the *U S* steel industry A vintage capital model

HIROTAKA TAKEUCHI, Ph.D California (Berkeley) 1977 Productivity analysis as a resource management tool in the retail trade

PATSY R. TARULLO, Ph D. Pittsburgh 1976. American Telephone & Telegraph Company. A survey of its development through basis strategy and structure.

GREGORY C. TASSEY, Ph.D George Washington 1978. A demand-shift model of economic performance in technology-based industries

GEORGE S. TEMPLE, Ph.D. Montana State 1978. A dynamic economic systems community impact model applied to coal development in the Northern Great Plains.

RICHARD C. TEPEL, Ph.D. Brown 1978. Optimizing behavior and efficiency of regional power pools.

WEN-LEE TING, Ph.D. New York 1977. Transfers of technology by multinational firms in the industrialization of Singapore.

RICHARD D. TRAINER, Ph.D Notre Dame 1977 A competitive alternative to rate of return regulation of common carriers: AT&T.

HELEN T. VEITH, Ph.D. Minnesota 1977. Public transportation in a circular city.

SARAH P. VOLL, Ph.D. New Hampshire 1978. Tech-

nological transfer in large-scale agricultural projects: The role of private enterprise.

BARRY R. WEINGAST, Ph.D. California Institute of Technology 1978. A representative legislature and regulatory agency capture.

STEVEN R. WEISBROD, Ph.D. Chicago 1978. The regulation of the securities portfolios of life insurance companies.

JOHN V. WELLS, Ph.D. Yale 1978. The origins of the computer industry: A case study in radical technological change.

LIONEL WILLIAMSON, Ph D Missouri (Columbia) 1977. The role of farmer cooperatives in small farm development

NEIL WRIGHT, Ph.D. Massachusetts Institute of Technology 1977 Four essays in industrial organization.

MOHAMMED B YUSOFF, Ph D Iowa State 1977 An econometric analysis of the world natural rubber industry.

Agriculture and Natural Resources

LEE J ALSTON, Ph.D Washington 1978 Costs of contracting and the decline of tenancy in the South, 1930 60.

RAPHAEL H AMIT, Ph D. Northwestern 1977. Petroleum reservoirs exploitation When and how

MOHD S BAKAR, Ph D Wisconsin (Madison) 1977 An economic analysis of supply response of rice acreage and management of rice reserves in peninsular Malaysia

TIMOTHY G BAKER, Ph.D Michigan State 1978 An economic and financial projections model of the *U S* farming sector.

ANTONIO L BANDEIRA, Ph D Purdue 1977 Capital-labor ratios in small rural firm-households in Brazil.

DAVID E. BANKER, Ph.D. Purdue 1977. A model for the analysis of alternative pricing policies in federal milk marketing orders.

KENNETH H BAUM, Ph.D Iowa State 1978 A national recursive simulation and linear programming model of some major crops in *U S* agriculture

HALIL C BEDESTENCI, Ph.D Ohio State 1978. Foreign market demand for *U S* soybeans and soybean products: A policy of approach.

RICHARD H. BERNSTEN, Ph D. Illinois 1977 Constraints to higher rice yields in the Philippines.

DAVID A. BESSLER, Ph D. California (Davis) 1977 Foresight and inductive reasoning: Analysis of expectations on economic variables with California field crop farmers.

TRAN BICH, Ph D. State University of New York (Binghamton) 1978. An econometric analysis of the role of air and water residuals in the production technology for steam-generating electric plants

DAVID M. BLITZER, Ph.D. Columbia 1978. An optimal control analysis of leasing federal oil.

ROBERT BOYNTON, Ph.D. Michigan State 1978. Performance of the *U.S.* dairy subsector as affected by the vertical coordination processes between co-operatives and proprietary handlers.

JON A. BRANDT, Ph.D. California (Davis) 1977. An economic analysis of the processing tomato industry.

PETER G. BUSHNELL, Ph.D. California (Davis) 1978. Dynamic analysis of the world almond market and the *U S* almond marketing order.

LEONARD A. CARLSON, Ph.D. Stanford 1977. The Dawes Act and the decline of Indian farming.

PRADIT CHARSOMBUTI, Ph.D. Illinois 1978. The redistribution of the labor force between the agricultural and nonagricultural sectors in Thailand

DENNIS C. CORY, Ph.D Iowa State 1977. Estimation of regional planted acreage and soil erosion losses for alternative export demand projections and conservation technologies: A macro-econometric approach.

JOHN A CRAVEN, Ph D. Illinois 1977 An econometric analysis of the *U S* feedgrain-livestock economy.

TIM DEYAK, Ph.D State University of New York (Binghamton) 1978 Generic congestion and relative rates of quality interference The case of recreational multiple use

RICHARD S DOWELL, Ph.D Chicago 1977 Risk diversification and land tenure in *U S* agriculture, 1890–1970

RICHARD DUNFORD, Ph D Wisconsin (Madison) 1977 An economic analysis of changes in the ownership and use of rural real estate in southwestern Wisconsin

BRIAN C. D'SILVA, Ph D Iowa State 1978. Factors affecting farmland ownership in Iowa

GEORGE M. EASTHAM, Ph.D. Claremont 1978. Toward measuring the demand for clean air A study in the theory and measurement of the demand for environmental quality.

EMMETT W ELAM, Ph.D Illinois 1978. A strong form test of the efficient market model applied to the *U S* hog futures market.

ELDON L ERICKSON, Ph.D. Iowa State 1978 A model for economic evaluation of alternative waste management systems in Iowa beef cattle feedlots.

KENNETH W. ERICKSON, Ph D. Wisconsin (Madison) 1978. Equity and efficiency of metallic mineral taxation policies in Wisconsin.

RAY K. ERICSON, Ph.D. Colorado State 1978. Water quality values in outdoor recreation.

MARK EVANS, Ph.D. New Mexico 1977. Evaluating the implications of structural change: A multiregional input-output model of the four corners states.

FRANK A. FENDER, Ph.D. Purdue 1978. Use of network techniques in the management of crop breeding research.

SILVIO J. FLAIM, Ph.D. Cornell 1978. Federal income taxation of the *U.S.* petroleum industry and the depletion of domestic reserves.

KLAUS K. FROHBERG, Ph.D. Illinois 1977. Optimal soil loss over time from a societal viewpoint.

DANIEL L. GALT, Ph.D. Cornell 1977. Economic weights for breeding selection indices: Empirical determination of the importance of various pests affecting tropical maize.

PHILIP GARCIA, Ph.D. Cornell 1978. Market linkage of small farms: A study of the maize market in northern Vera Cruz, Mexico.

GENE A. GERMAN, Ph.D. Cornell 1978. Dynamics of food retailing, 1900–75.

REZA M. GHANBARI, Ph.D. Michigan State 1977. Optimal allocation of a renewable resource: A bioeconomic model of the Great Lakes Whitefish Factory.

CHRISTINA H. GLADWIN, Ph.D. Stanford 1977. A model of farmers' decisions to adopt the recommendations of Plan Puebla.

ERIC K. GREEN, Ph.D. Kentucky 1977. Economic efficiency of three synthetic fuel processes.

HECTOR E. GONZALEZ-MENDEZ, Ph.D. Iowa State 1977. Grain marketing gains in Iowa and the use of price forecasting models: A Bayesian decision approach.

RAYMOND S. HARTMAN, Ph.D. Massachusetts Institute of Technology 1977. An econometric simulation model of the *U S.* copper industry.

JULIO HERNANDEZ, Ph.D. North Carolina State 1978. Estimation of demand parameters for eggs.

SHIN H. HUH, Ph.D. Minnesota 1978. The preventive and incidental demands for pesticides· An economic analysis of the demand for herbicides and insecticides used by selected corn producers in Minnesota.

TERRY L. HUSKEY, Ph.D. Washington (St. Louis) 1977. Park characteristics and the demand for recreational trips.

FRANCOIS KAMAJOU, Ph.D. Illinois 1978. Government financing of the development of small farm agriculture in the center-south province of Cameroon.

JAMES KASAL, Ph.D. Colorado State 1978. The impacts of constrained soil loss, fertilizer use, and land use diversity on farm income.

RAHMAN KHOSHAKHLAGH, Ph.D. New Mexico 1977. Forecasting the value of water rights: A case study of New Mexico.

BILL H. KINSEY, Ph.D. Stanford 1977. Agricultural technology and rural development in the rainfed maize area of southeastern Zambia.

KHAISRI KONJING, Ph.D. Minnesota 1977. An analysis of the economic performance of the *U.S.* corn futures market.

RAY KOPP, Ph.D. State University of New York (Binghamton) 1978. The implications of pollution and pollution control equipment for productive efficiency in the electric power industry.

WILLIAM H. LESSER, Ph.D. Wisconsin (Madison) 1977. Marketing systems for warm water aquaculture species in the upper Midwest.

ARMANDO A. LLOP, Ph.D. California (Davis) 1978. Economics of irrigation under salinity conditions: The case of Mendoza, Argentina.

IGNEZ G. LOPES, Ph.D. Purdue 1977. Time allocation of low-income Brazilian households: A multiple job holding model.

MAURO D. LOPES, Ph.D Purdue 1977. Alternative fiscal means to mobilize resources from agriculture: The case of Brazil.

ERNST LUTZ, Ph.D. California (Berkeley) 1977. Grain reserves and international price stabilization.

JOHN K. LYNAM, Ph.D. Stanford 1978. An analysis of population growth, technical change, and risk in peasant, semiarid farming systems. A case study of Machakos District, Kenya.

ANDREW M. MCGREGOR, Ph D. Cornell 1978. The Lome Convention and the *ACP* sugar exporters· The political economy of conflicting policies.

MICHAEL V. MARTIN, Ph.D. Minnesota 1978 An economic analysis of the social cost of regulated value-of-service wheat and barley rail rates in the upper Midwest

RONALD L. MEEKHOF, Ph.D. Michigan State 1978. The economic feasibility of utilizing waste heat from electrical power plants in integrated agricultural and aquacultural systems under Michigan conditions.

BRYAN E. MELTON, Ph.D. Iowa State 1977. An economic analysis of concentrate vs. roughage feeding for finishing beef steers.

WILLIAM H. MEYERS, Ph.D. Minnesota 1977. Long-run income growth and world grain demand: An econometric analysis.

DWIGHT D. MINAMI, Ph.D. California (Davis) 1977 The economic analysis of market control in the California cling peach industry.

HENDRIK C. MOLSTER, Ph.D. Stanford 1978. Methods of estimating fertilizer response with an application to area use in Jogjakarta, Indonesia.

JOHN L. MORRIS, Ph.D. Cornell 1978. An economics analysis of cyclical variations in the *U.S.* beef industry.

BERNARD J. MORZUCH, Ph.D. Missouri (Columbia) 1977. Technology and climatic effects in aggregate functions.

KOOSWARDHONO MUDIKDJO, Ph.D. Wisconsin (Madison) 1978. Changes in agrarian production

structure under an agrarian reform structure: Chile 1965–70.

UPALI NANAYAKKARA, Ph.D. Michigan State 1977. The economics of country fairground use and the potentials for profitable future operations through use expansion: A case study of the fairgrounds project in Emmet County, Michigan.

MACK C. NELSON, Ph.D. Illinois 1978. An economic analysis of the long-run productivity impacts of soil erosion control.

ROBERT C. OELHAF, Ph.D. Maryland 1976. The economics of organic farming.

CHARLES E. OVERTON, Ph.D. Purdue 1977. A model for long-run feed ingredient procurement planning under risk.

ANTONIO C. PINHEIRO, Ph.D. Iowa State 1978. Corn supply and water-nitrogen demand functions based on experimental data: The uncertainty case.

CHIRMSAK PINTHONG, Ph.D. Stanford 1977. A price analysis of the Thai rice marketing system.

NIMAL F. RANAWEERA, Ph.D. California (Davis) 1978. Land settlement in Sri Lanka: The Mahaweli Ganga Project.

JOHN A. REEDER, Ph.D. State University of New York (Buffalo) 1978. Corporate social involvement at the local level: A study of social responsibility in San Antonio, Texas.

ROLAND L. REYNOLDS, Ph.D. Washington State 1977. Interrelationships between market structure and pollution abatement activities.

STEVEN K. RIGGINS, Ph.D. Cornell 1978. Corn marketing in western New York.

JAMES F ROACH, Ph.D. New Mexico 1977. An economic model for the Rio Grande Drainage Basin, New Mexico.

GIL R. RODRIGUEZ, Ph.D. Purdue 1978. The consideration of risk in agricultural policies: The Philippine experience.

JEFFREY S. ROYER, Ph.D. Iowa State 1978. A general non-linear programming model of a producers cooperative association in the short run.

EUGENIA M. RUBINSTEIN, Ph.D. Minnesota 1977. The economics of foot and mouth disease control and its associated externalities.

FRED M. RUKANDEMA, Ph.D. Cornell 1978. Resource availability, utilization, and productivity on small-scale farms in Kakamega District, western Kenya.

WILLIAM H. SANDER, Ph.D. Cornell 1978. Some economic, institutional, and political aspects of irrigation planning and development: Lessons from the Bureau of Reclamation experience.

ORHAN SAYGIDEGAR, Ph.D. Iowa State 1977. Analysis of interaction between soil conservation and agricultural production in the United States using a multigoal linear programming model.

ALI SHAMS, Ph.D. Southern Illinois 1977. The feasibility of solar house heating: A study in applied economics.

BASIL M. H. SHARP, Ph.D. Wisconsin (Madison) 1978. The economics of managing water quality: A multiobjective analysis of alternative policies.

ROBIN N. SHAW, Ph.D. Cornell 1977. Universal product code scanning systems: The retail experience 1974–76.

WILLIAM SIMS, Ph.D. Toronto 1978. The economics of sewer affluent charges.

JAMES L. SMITH, Ph.D. Harvard 1977. Bidding behavior for offshore petroleum leases.

NANCY M. SNYDER, Ph.D. Southern Illinois 1977. The relative income distribution effects of energy price changes.

AUGUSTO C. SOARES, Ph.D. Ohio State 1977. Resource allocation and choice of enterprise under risk on cotton farms in northeast Brazil.

JOHN D. SPRIGGS, Ph.D. Minnesota 1977. An econometric analysis of the factors affecting Australia's grain exports.

RUANGDEJ SRIVARDHANA, Ph.D. Colorado State 1977. Water uses and an optimizing model of water pollution.

JEFFREY C. STIER, Ph.D. Wisconsin (Madison) 1977. The economics of a dual externality: Agriculture and Canada geese in Wisconsin.

EDWARD C. THOR, Ph.D. California (Berkeley) 1978. An economic framework for wildland planning decision making.

WALLACE E. TYNER, Ph.D. Maryland 1977. Energy resource development and economic development in India.

FAHRI M. UNSAL, Ph.D. Cornell 1978. The impact of an alternative price policy which incorporates transportation and storage costs into the set of *TMO* purchase and sale prices with Kenya as a base.

PETER J. VAN BLOKLAND, Ph.D. Illinois 1977. The economic consequences of hail suppression and demand in 1985 on foodgrains, feedgrains, and oilmeals in the United States.

DAVID J. WALKER, Ph.D. Iowa State 1977. An economic analysis of alternative environmental and resource policies for controlling soil losses and sedimentation from agriculture.

RODNEY L. WALKER, Ph.D. Purdue 1978. Short-run policies for the *U.S.* grain and livestock sectors: An application of control theory.

FRANK A. WARD, Ph.D. Colorado State 1977. The welfare effects of a market allocation of an exhaustible resource.

MICHAEL K. WOHLGENANT, Ph.D. California (Davis) 1978. An economic analysis of the dynamics of price determination: A study of the California grape-wine industry.

CHARLES L. WRIGHT, Ph.D. Ohio State 1977. The economics of grain transportation and storage: A Brazilian case study.

MING WU WU, Ph.D. Michigan State 1978. Supply response and production outlook for tart cherries in Michigan.

Manpower, Labor, and Population; including Trade Unions and Collective Bargaining

JOHN ABOWD, Ph.D. Chicago 1977. An econometric model of the U.S. market for higher education

STEVEN G. ALLEN, Ph.D. Harvard 1978. Absenteeism and the labor market.

NEIL ALPER, Ph.D. Pittsburgh 1977. The impact of the cognitive and noncognitive aspects of education on skilled workers: A case study.

JOSEPH M. ANDERSON, Ph.D. Harvard 1977. An economic-demographic model of the U.S. labor market

WILLIAM C. APGAR, JR., Ph.D. Harvard 1978. Occupational, industrial, and geographical mobility. A human capital approach

AYSE ARITURK, Ph.D. Johns Hopkins 1978. General equilibrium effects of Turkish migration

ABDOLHI ARMAND, Ph.D. Maryland 1976 The rate of return to education in Iran

HAROLD E. BANGUERO, Ph.D. North Carolina (Chapel Hill) 1977. The social and economic determinants of fertility in Columbia

JUDITH BANISTER, Ph.D. Stanford 1977. The current vital rates and population size of the People's Republic of China and its provinces

ROBERTA BARNES, Ph.D. Michigan 1977. Household composition effects on household expenditure patterns.

HENRY BARTEL, Ph.D. Indiana 1978. An economic model of the short-run demand for higher education in Indiana: A human capital approach.

MOHAMMAD M. BEHKISH, Ph.D. Indiana 1977. Economics of investing in human capital: The case of Iran.

CHRIS J. BERGER, Ph.D. Wisconsin (Madison) 1978. An examination of operant conditioning and expectancy theory accounts of instrumental work behavior and satisfaction.

YVES R. BIZIEN, Ph.D. Tufts 1977. Population and economic development.

ARTHUR E. BLAKEMORE, Ph.D. Southern Illinois 1977. The dynamics of unionization, the industrial structure, and wage inflation.

DAVID J. BOWEN, Ph.D. California (Berkeley) 1977. Teacher collective bargaining and multilateralism.

RICHARD A. BRADLEY, Ph.D. California (Riverside) 1977. Some economic aspects of fertility behavior.

ANDREW W. BRAUNSTEIN, Ph.D. Rutgers 1978. Labor supply behavior of female heads of households: Extending the classical model to a dynamic framework.

CHARLES H. BREEDEN, Ph.D. Virginia Polytechnic Institute 1977. Third-party effects and labor entitlements: An economic perspective.

BILLIE ANN BROTMAN, Ph.D. Notre Dame 1978. The impact of the *NLRB*'s deferral policy on unfair labor dispute settlements involving unilateral changes in working conditions and discriminatory practices.

WADSWORTH S. CAUCHOIS, JR., Ph.D. California (Davis) 1977. Federal manpower programs in San Joaquin County, 1963–73: Magnitudes and impacts.

STEPHEN CHAIKIND, Ph.D. City (New York) 1978. The quality adjusted demand for public elementary school teachers A labor market analysis.

CLARENCE M. CONDON III, Ph.D. South Carolina 1978. The effects of achievement motivation and socioeconomic background on the earnings of young men.

MARGUERITE M. CONNERTON, Ph.D. Harvard 1978. Accident control through regulation: The 1969 Coal Mine Health and Safety Act experience

JOSEPH CORDES, Ph.D. Wisconsin (Madison) 1977 The economics of "actual" compensation A theoretical and empirical perspective.

ROBERT F. COTTERMAN, Ph.D. Chicago 1978 A theoretical and empirical analysis of the labor supply of "older" males

JOHN W. DANSBY, Ph.D. Kentucky 1976. Inventories and orders in a model of industrial layoffs.

LEIF DANZINGER, Ph.D. Yale 1978 The theory of optimal labor contracts.

THOMAS F. DAVIS, Ph.D. Pittsburgh 1976. Measuring the convergence of incomes and its affect upon fertility decisions. The American experience, 1947–73.

CARL J. DEMERY, Ph.D. Wayne State 1977. An estimate of the empirical relationship between urban labor force participation, industrial specialization, unemployment rates, and city size.

BARBARA DEVANEY, Ph.D. Michigan 1977 The labor supply of married women: An analysis of the allocation of time to market and nonmarket activities.

JOHN A. DIXON, Ph.D. Harvard 1978. Economic aspects of rural to rural migration and land settlement in East Asia.

MARTIN DOOLEY, Ph.D. Wisconsin (Madison) 1977. An analysis of the labor supply and fertility of married women with grouped data from the 1970 U.S. Census.

JAMES E. DUGGAN, JR., Ph.D. Pennsylvania State 1977. A time-series analysis of secondary labor force participation.

DAVID A. DUMONT, Ph.D. State University of New York (Albany) 1977. Employment improvement, human capital, and secondary labor market aspects

of the Manpower Training Program in South Carolina.

RANDALL W. EBERTS, Ph.D. Northwestern 1978. An economic analysis of municipal governments in a metropolitan setting.

LEE E. EDLEFSEN, Ph.D. Harvard 1978. The joint determination of the numbers, timing, and spacing of children.

RICHARD ELLSON, Ph.D. Florida 1977. The impact of metropolitan consolidation on fiscally induced migration: An econometric simulation approach

JOHN C. EVANS, Ph.D. Chicago 1978 The social opportunity cost of labor in Canada.

HENRY S FARBER, Ph.D. Princeton 1977 The United Mine Workers and the demand for coal: An econometric analysis of union behavior.

SHAHROKH FARDOUST, Ph D. Pennsylvania 1978. Risk-taking behavior, socioeconomic background, and distribution of income. A theoretical and empirical analysis.

JACK FRISCH, Ph.D. Princeton 1978. The interstate variation in workers compensation A neoclassical and a radical explanation

CAROL A. GAMBILL, Ph.D Missouri (Columbia) 1978 A conceptual framework for a comprehensive labor market information system

ROBERT GITTER, Ph.D Wisconsin (Madison) 1978. A simultaneous equation model of the labor force participation rate of prime age males.

ITZHAK GOLDBERG, Ph.D Chicago 1977. Enforcement of work discipline: An economic analysis.

AMYRA GROSSBARD, Ph D Chicago 1978 The economics of polygamy

ROBERT P. HAGEMANN, Ph.D. Florida State 1977. An economic demographic regression model for projecting K-12 public school enrollments: The case of Florida.

JOAN HANNON, Ph D. Wisconsin (Madison) 1977 The immigrant worker in the promised land: Human capital and ethnic discrimination in the Michigan labor market, 1888-90.

CLIFFORD B. HAWLEY III, Ph.D. Duke 1977. The effects of labor market distortions on real wages. A general equilibrium analysis.

MARTHA HILL, Ph.D. Michigan 1977. The decision by young adults to split off from their parents' households.

CHIRA HONGLADAROM, Ph.D. Washington 1978. The effect of child mortality on fertility in Thailand.

JOHN J. HOOVER, Ph.D. Notre Dame 1978. The impact of industrial organization on industrial relations as applied to conglomerate corporations and coalition bargaining.

ROBERT N. HORN, Ph.D. New Hampshire 1978. Labor market segmentation in New England: Empirical and case studies.

MARY J. HORNEY, Ph.D. Duke 1977. Household decision making: A game-theoretic approach.

CHUN-YANG HSU, Ph.D. North Carolina State 1977. Education, production, and labor substitution in agriculture.

JOHN F. HULPKE, California (Berkeley) 1977. Equal employment opportunity/affirmative action programs in banks: Why some programs succeed where others fail.

LOUIS S. JACOBSON, Ph.D. Northwestern 1977. Earnings losses and worker displacement when employment declines in the steel industry.

THOMAS M. KICKHAM, Ph.D. Pittsburgh 1977. A micro-economic analysis of the relationship between American savings and fertility. 1950-70.

RANDALL H. KING, Ph.D. Ohio State 1978. The labor market consequences of dropping out of high school

DAVID R KNOWLES, Ph.D. Washington State 1978. Evaluating the impact of federal counterrecessional job creation programs within the state of Washington

WILLIAM KRUSE, Ph.D. State University of New York (Binghamton) 1978. Earnings of young males: A human capital approach

IRA T. LAPIDES, Ph.D. Tennessee (Knoxville) 1978. Long-run returns from vocational training and the "Screening Hypothesis": A follow-up study.

JOANNE LINNEROOTH, Ph D. Maryland 1977. The evaluation of public programs affecting population mortality.

BARRY MCCORMICK, Ph.D. Massachusetts Institute of Technology 1977 Aspects of labor markets: Three essays.

M. BRIAN MCDONALD, Ph.D. Pennsylvania 1978. The distribution and production of education: A disaggregate analysis

INDRA MAKHIJA, Ph D. Chicago 1977. The economic contribution of children and its effects on fertility and schooling. Rural India.

DAVID H. MALMQUIST, Ph.D City (New York) 1978. A lagged adjustment model of foreign worker migration to the Federal Republic of Germany.

STEVEN MANASTER, Ph.D. Chicago 1977. Wage adjustments to anticipated and unanticipated inflation.

FRANK MARTIN, Ph.D. Tulane 1977. Occupational redistribution as a source of economic growth.

LINDA G. MARTIN, Ph.D. Princeton 1978. Measuring completeness of death registration in destabilized populations.

ARTHUR C. MEAD, Ph.D. Boston College 1978. Migration and employment growth in nonmetropolitan areas of the United States: 1960-70.

FATHOLLA MIZRA-BAGHERI, Ph.D Pennsylvania 1978. Distributional impacts of macro-economic fluctua-

tion on the structure of earnings: A longitudinal approach.

DAVID A. MOLNAR, Ph.D. Harvard 1977. Regional wage differentials: A multisector analysis.

MARK J. MORLOCK, JR., Ph.D. Washington State 1978. Public service employment as a manpower policy.

RANDALL OLSEN, Ph.D. Chicago 1977. An econometric analysis of family labor supply.

LYNN PARINGER, Ph.D. Wisconsin 1978. Determinants of work loss and medical care utilization for specific illnesses.

JEFFREY M. PERLOFF, Ph.D. Massachusetts Institute of Technology 1977. The wage change process in the construction industry.

COLLIS PHILLIPS, Ph.D. Syracuse 1977. An econometric analysis of the income distribution effects of the Manpower Development and Training Act.

LISE POULIN-SIMON, Ph.D. McGill 1977. Le loisir industriel et le chômage au Canada: Une histoire économique.

JOHN RAISIAN, Ph.D. California (Los Angeles) 1978. Cyclic variations in hours, weeks, and wages.

JONATHAN B. RATNER, Ph.D. Yale 1978. Costs of labor adjustment and the demand for labor: A case study in the microfoundations of macroeconomics.

MICHELLE RIBOUD, Ph.D. Chicago 1977. An analysis of earnings distribution in France.

W. CRAIG RIDDELL, Queen's (Kingston) 1977. The determinants of negotiated wage changes in Canadian industry, 1953-73: A study based on wage contract data.

MARIO RIZZO, Ph.D. Chicago 1977. Rents, property values, and the cost of crime to victims.

CHRISTOPHER ROBINSON, Ph.D. Chicago 1977. Allocation of time across the day: An analysis of the demand and supply of shiftworkers.

JACK L. RODGERS, Ph.D. Minnesota 1977. The effects of college quality on earnings.

ERNESTO Q. RODRIGUEZ, Ph.D. Pittsburgh 1976. Interstate labor force migration in Mexico: 1960-70.

ROBERT A. ROSENTHAL, Ph.D. Boston 1978. Obstacles to the diffusion of worker participation.

RUSSELL T. ROSS, Ph.D. Duke 1977. Female labor supply in New Zealand.

STEPHEN A. RUBENFELD, Ph.D. Wisconsin (Madison) 1977. Urban transit: Public ownership and collective bargaining.

ROBERT S. RYCROFT, Ph.D. Maryland 1978. Aspects of the labor market for young men: Labor supply and unemployment.

ANDRÉ SAPIR, Ph.D. Johns Hopkins 1978. External migration in a growing labor-managed economy: The case of Yugoslavia.

SANDRA SCHICKELE, Ph.D. Chicago 1977. The social opportunity cost of urban labor in the United States.

ROBERT H. SCHNORBUS, Ph.D. Case Western Reserve 1978. Labor market segmentation: A study of barriers to occupational mobility.

ARTHUR SCHWARTZ, Ph.D. Michigan 1978. The effect of benefits and overtime costs on the short-run cyclical demand for labor in the automobile industry in Michigan.

VED P. SHARMA, Ph.D. Washington (St. Louis) 1977. An economic analysis of fertility behavior in India.

RICHARD G. SHEEHAN, Ph.D. Boston College 1978. Wage-price controls and the aggregate real wage.

DAVIMDER SINGH, Ph.D. South Carolina 1978. Wage determination in *U.S.* manufacturing industries, 1958-72: A collective bargaining approach.

RICHARD H. STECKEL, Ph.D. Chicago 1977. The economics of *U.S.* slave and southern white fertility.

SUDARSONO, Ph.D. Tennessee (Knoxsville) 1977. The degree of responsiveness of the Tennessee labor force, 1970: A cross-sectional analysis by county.

DANIEL SUMNER, Ph.D. Chicago 1978. Off-farm labor supply and earnings of farm family members.

GEORGE TAUCHEN, Ph.D. Minnesota 1978. The minimum wage and job search.

NIGEL TOMES, Ph.D. Chicago 1978. A model of child endowments, and the quality and quantity of children.

MAI-TRANG THI NHU TRAN, Ph.D. Southern Illinois 1978. The derivation and analysis of human resource proxies utilizing principal component analysis.

JOHN G. TREBLE, Ph.D. Northwestern 1978. On the theory of interregional migration.

JOHN A. TURNER, Ph.D. Chicago 1977. The effect of fertility on the distribution of earnings.

JOE U. UMO, Ph.D. Indiana 1978. Economics of human capital production and investment: The case of Nigerian higher education.

ALEJANDRO VELEZ, Ph.D. Florida 1977. An economic analysis of the fertility of male-headed Spanish origin households in the United States in 1970.

RICHARD VICTOR, Ph.D. Michigan 1977. The effects of municipal unionism on wages and employment.

DAVID G. WAGAMAN, Ph.D. Nebraska (Lincoln) 1977. Public employee impasse resolution: A historical examination of the Nebraska experience with some comparisons to the New York experience.

CHARLES P. WEBER, Ph.D. Wayne State 1977. A survey and analysis of wage determination for public employees in the Detroit, Michigan area.

GREGORY C. WEEKS, Ph.D. Washington State 1977 The dual labor market, the Phillips curve, and th class conflict business cycle: A synthesis.

STEPHEN D. WILSON, Ph.D. West Virginia 197

Optimization of Army enlisted accessions through variable rates of pay.

PETER WRAGE, Ph.D. Indiana 1978. A human capital approach to regional earnings disparities and the ameliorating effects of labor migration.

STEPHEN L. ZABOR, Ph.D. Northwestern 1977. The measurement and analysis of executive compensation.

Welfare Programs; Consumer Economics; Urban and Regional Economics

JEFFREY L. ADAMS, Ph.D. Pittsburgh 1976. The costs of and benefits from reducing hospital employee turnover.

JAMES R. ANDERSON, Ph.D. Florida State 1978 The economic value of graduate study sponsored by the U.S. Air Force.

RUTH ANDRESS, Ph.D. South Carolina 1977. The role of attitudes in the labor force participation of married women

NOLAN H. BARKER, JR., Ph.D. Virginia Polytechnic Institute 1977. The federal food stamp and related in-kind commodity distributions: Economic history and evaluation in a public choice perspective.

JUDITH D. BENTKOVER, Ph.D. Tufts 1978 An empirical investigation of the impact of inflation and unemployment on public vis-à-vis private community hospital utilization

PETER H. BOHLING, Ph.D. Massachusetts 1977. An econometric model for determining poverty rates.

KATHERINE L. BRADBURY, Ph.D. Massachusetts Institute of Technology 1977. Housing supply in a metropolitan area.

JOSEF M BRODER, Ph.D. Michigan State 1977. The provision of court services An inquiry into the allocation of opportunities to rural communities.

JOHN L. BUNGUM, Ph.D. Nebraska (Lincoln) 1977 A study of selector aspects of textbooks and textbook writers in international economics, 1945–74.

BURKE K. BURRIGHT, Ph.D. California (Los Angeles) 1977 Cities and travel: An aggregate, equilibrium model of urban travel volumes, traffic congestion, and land area.

STEVEN T CALL, Ph.D. Indiana 1977. The sewage treatment service: Economies of scale and optimal industrial surcharge strategies.

GERALD A. CARLINO, Ph.D. Pittsburgh 1976. Agglomeration of manufacturing activity in metropolitan areas: Theory and measurement.

WILLIAM J. CARROLL, Ph.D. Pennsylvania State 1977. Measuring the impact of government activity using residential housing values: A comparison of time-series and cross-sectional approaches.

KARL E. CASE II, Ph.D. Harvard 1977. Intrajurisdictional variation in effective rates of property taxation.

CONSTANTINOS A. CHRISTOFIDES, Ph.D. Lehigh 1977. An econometric model for the northeastern Pennsylvania region.

HOPE CORMAN, Ph.D. City (New York) 1978. The effects of apprehension, conviction, and incarceration on crime in New York State.

HENRY CROUCH, Ph.D. South Carolina 1977. Three essays on wage differentials between white males and females.

DONALD J. CYMROT, Ph.D. Brown 1978. An economic analysis of private pensions.

GESTUR B. DAVIDSON, Ph.D. Minnesota 1978. Manpower substitutions and hospital efficiency.

DAVID M. DE FERRANTI, Ph.D. Princeton 1978. Tests of seven hypotheses on welfare dependency and family disintegration.

DOUGLAS B. DIAMOND, Ph.D. Chicago 1978. Income and residential location in urban areas.

KEITH D. EDWARDS, Ph.D. Pittsburgh 1977. Cost benefit implications of a city day care program.

SENA EKEN, Ph.D. Pittsburgh 1976. General equilibrium analysis of local taxes and expenditures: A case study of the Pittsburgh *SMSA*.

KATHLEEN F. FELDSTEIN, Ph.D. Massachusetts Institute of Technology 1977. The economics of public libraries.

BARRY FISHMAN, Ph.D. California (Los Angeles) 1977. A simple robust test of the empirical relevance of local public services and neighborhood characteristics in the choice of residential location.

DIANE FLAHERTY, Ph.D. New York 1978. Decentralization: Workers' control and regional income differences in Yugoslavia.

JACOB GESTHALTER, Ph.D. City (New York) 1978. Economic analysis of mother's education on the child's health and the demand for health inputs.

FRANK D. GIANFRANCESCO, Ph.D. Maryland 1976. Insurance and medical care expenditure: A theoretical analysis of market efficiency.

JOHN M. GRANA, Ph.D. Massachusetts 1978. The impact of reimbursement mechanisms on management decision making in the long-term health care industry in Massachusetts.

DIANNE GREEN, Ph.D. Pittsburgh 1977. Some effects of social security programs on income distribution in Costa Rica.

TIMOTHY J. GRONBERG, Ph.D. Northwestern 1978. The interaction of markets in housing and local public goods: A simultaneous equations approach.

RONALD J. GUNDERSON, Ph.D. Nebraska (Lincoln) 1977. The determination of structural differences influencing the growth of the rural Old West region.

HAMILTON W. HELMER, Ph.D. Yale 1978. The Vermont economy, 1950–70: A study of regional growth.

ROBERT J. HIGHSMITH, Ph.D. Indiana 1978. An

economic inquiry into the production and distribution of police services.

SAUL HOFFMAN, Ph.D. Michigan 1977. Discrimination over the life cycle: A longitudinal analysis of black-white experience-earnings profiles.

JOHN M. HORNE, Ph.D. Carleton 1978. Copayment and utilization of publicly insured hospital services in Saskatchewan: An empirical analysis.

ALLEN HYMAN, Ph.D. California (Los Angeles) 1978. Law and economics of local government annexation and boundary change.

KEITH R. IHLANFELDT, Ph.D. Washington (St. Louis) 1978. Moving costs and the demand for housing.

STEPHEN S. IRVINE, Ph.D. Tennessee (Knoxville) 1978 A study of state per capita income differentials with special reference to South–non-South differentials, 1950, 1960, and 1970.

SUSAN S. JACOBS, Ph.D. Brown 1978. A theoretical analysis of airport location and negative externality Optimality, efficiency, and equity.

MANUEL H. JOHNSON, Ph.D. Florida State 1977 An assessment of the impact of nuclear power plant construction and operation on small regions.

GREGORY KILGARIFF, Ph.D. Notre Dame 1977. Impact of inflation on urban areas in the state of New York

RICKEY C KIRKPATRICK, Ph D Tulane 1978. Sources of regional comparative advantage in *U.S* manufacturing industries.

BRIAN LEBOWITZ, Ph D Yale 1977. An economic approach to criminal sentencing.

KEITH B. LEFFLER, Ph.D California (Los Angeles) 1977. An economic analysis of physician licensure

WILHELMINA LEIGH, Ph.D. Johns Hopkins 1978 An analysis of the growth of the stock of housing as an indicator of the growth of housing services in the United States, 1950 70.

PETER LINNEMAN, Ph.D. Chicago 1977. An analysis of the demand for residence site characteristics.

BARRY A. LOVE, Ph.D. Virginia 1978 An economic evaluation of the Food Stamp Program.

JAMES C. LYNCH, Ph.D. Utah 1978. Income generating crimes· An economic approach to crime causation.

BRUCE F. MCELROY, Ph D. California (Berkeley) 1978. Unit pricing six years after introduction: An analysis and an extension.

WILLIAM J. MCGUIRE, Ph.D. North Carolina (Chapel Hill) 1978. An economic model of criminal correctional institutions: Empirical cost analysis and cost benefit measures.

WILLIAM MCNAUGHT III, Ph.D. Harvard 1978. The simulation of search behavior in urban housing markets.

ROSS A. MARCOU, Ph.D. Princeton 1977. Studies of community homogeneity.

STEPHEN MARGOLIS, Ph.D. California (Los Angeles) 1978. Depreciation of capital in housing.

RASOOL MASOOMIAN, Ph.D. State University of New York (Binghamton) 1978. Housing demand: A case study of Tehran.

JOHN MULLEN, Ph.D. State University of New York (Binghamton) 1978 Consequences of replacing implicit subsidies with direct student grants on benefit incidence and enrollment in higher education.

VALERIE NELSON, Ph.D Yale 1977. Public provision and private markets A case study of vocational education.

LAURA NOWAK, Ph.D. City (New York) 1978. An evaluation of the Federal-State Rehabilitation Program in New Jersey as a means of income replacement

MEHMET A. ODEKON, Ph.D. State University of New York (Albany) 1977 The impact of education on the size distribution of earnings in Turkey

CLINTON V OSTER, JR, Ph.D Harvard 1978. The structure of intraurban household travel

A MEAD OVER, JR., Ph D Wisconsin (Madison) 1978 On the estimation of the primary care production function where output is an unobservable variable

GEORGI PALUMBO, Ph.D. Syracuse 1977 The impact of public employment on employment levels in major metropolitan areas

ITSCHAK D. PARIS, Ph D. Brown 1978 The distribution of city sizes Some aspects of mobility and change.

IAN C PARKER, Ph D Yale 1977 Studies in the economics of education.

MAHESH K PODAR, Ph.D. Pennsylvania State 1978 A benefit-cost analysis of fuel economy standards for passenger cars· Model years, 1981 84.

WILLIAM R POKROSS, Ph D Pittsburgh 1977 The occupational participation of younger black men. Variation among and recent changes in the larger metropolitan areas.

WILLIAM E. ROSEN, Ph D California (Davis) 1978. Externalities in central city and suburban crime. A simultaneous crime supply-police expenditure model

LOUIS F. ROSSITER, Ph.D North Carolina (Chapel Hill) 1977. A transcendental production function for health services. The community pharmacy.

ROBERT F. SCHLACK, Ph.D. Wayne State 1977. Residential land use patterns and local public policies· A study of the Detroit metropolitan area.

MAX M. SCHREIBER, Ph.D. South Carolina 1978. The development of the southern United States A test for regional convergence and homogeneity.

JAMES A. SCHUTTINGA, Ph.D. Massachusetts Institute of Technology 1977. Three empirical essays on the economics of hospital care.

BRUCE A. SEAMAN, Ph.D. Chicago 1978. A positive analysis of government cultural subsidies.

ROBERT A. SHAKOTKO, Ph.D. Minnesota 1977. Health, unobservable variables, and household decision making.

ROBERT SHEEHY, Ph.D. Florida 1977. An econometric analysis of consumer demand in Florida.

JOHN M. SIMPSON, Ph.D. Minnesota 1978. The charity market: An analysis of United Way.

SHELDON STEIN, Ph.D. Johns Hopkins 1978. Social security, interest rates, and life cycle consumption.

ROBERT L. STOUT, Ph.D. Pennsylvania State 1978. An economic analysis of the relationship of educational expenditures and migration in Pennsylvania counties, 1960-70.

VELMA THOMPSON, Ph.D. California (Los Angeles) 1978. Inefficient labor market discrimination under competitive conditions.

CRAIG THORNTON, Ph.D. Johns Hopkins 1978. Zoning, apartments, and land use interactions.

EUGENIA L. TOMA, Ph.D. Virginia Polytechnic Institute 1977 The economic organization of public education in the United States

GORDON R. TUSH, Ph D Nebraska (Lincoln) 1977. Effects of clients' socioeconomic characteristics and institutional variables on utilization of community mental health care. A case study.

ROLAND E. UBOGU, Ph.D. State University of New York (Albany) 1978. Highway congestion, work staggering, optimal tolls, and residential location: Three essays in transportation and urban economics.

BULENT UYAR, Ph.D. Pittsburgh 1977. Industrial-occupational employment projections: Pittsburgh *SMSA*.

JOSEPH J. VALENZA, Ph.D. George Washington 1977. Residential rehabilitation.

ROGER J. VAUGHAN, Ph.D. Chicago 1977. The value of urban open space

MICHAEL VERNARELLI, Ph.D. State University of New York (Binghamton) 1978. Locational distortion and black ghetto expansion.

ALFRED J. WATKINS, Ph.D. New School 1978. Uneven development in the *U S* system of cities.

CAROLYN L. WEAVER, Ph D. Virginia Polytechnic Institute 1977. The emergence, growth, and redirection of social security: An interpretive history from a public choice perspective.

JACK L. WERNER, Ph.D. Pennsylvania State 1977. The effects of insurance on hospital utilization and cost. A simultaneous equations model.

STEVEN P. WILKINSON, Ph.D. Southern Illinois 1977. The effect of state and federal funding on equality of education expenditure.

ANNOUNCEMENT

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1979 issue of the *Review* will carry the seventy-sixth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1979. This announcement is an invitation to send us information for the preparation of the list. This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1979 issue the names of those who will have been awarded the doctoral degree since June 1978 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3 x 5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ______

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ______

Degree Conferred (Ph.D. or D.B.A.) ______ Year ______

Dissertation Title: ______

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit educational organization, located at 1313 21st Avenue So., Nashville, Davidson County, Tennessee 37212. The Managing Editor is Professor George H. Borts, *American Economic Review*, Brown University, Robinson Hall, Providence, R.I. 02912. During the preceding 12 months the average number of copies printed for each issue was 27,765, the average paid circulation, 24,498; the average free distribution, 390; the average number of copies distributed, 24,888. Corresponding figures for the last issue before filing: 28,000 total number copies printed; 23,241 total paid circulation; 393 free copies distributed; 23,631 total distribution.